Leveraging Machine Learning and Molecular Dynamic Simulation to Explore IFRD1 Associated SNPs in Spinocerebellar Ataxia Type 18

Submitted by
Roman Bibi
62-FOC/MSBI/F22

Supervised by
Dr. Attiya Kanwal

Co-Supervised by
Dr. Noreen Akhtar

Department of Bioinformatics
Faculty of Computing & Information Technology
International Islamic University Islamabad

2025

Leveraging Machine Learning and Molecular Dynamic Simulation to Explore IFRD1 Associated SNPs in Spinocerebellar Ataxia Type 18

Submitted by
Roman Bibi
62-FOC/MSBI/F22

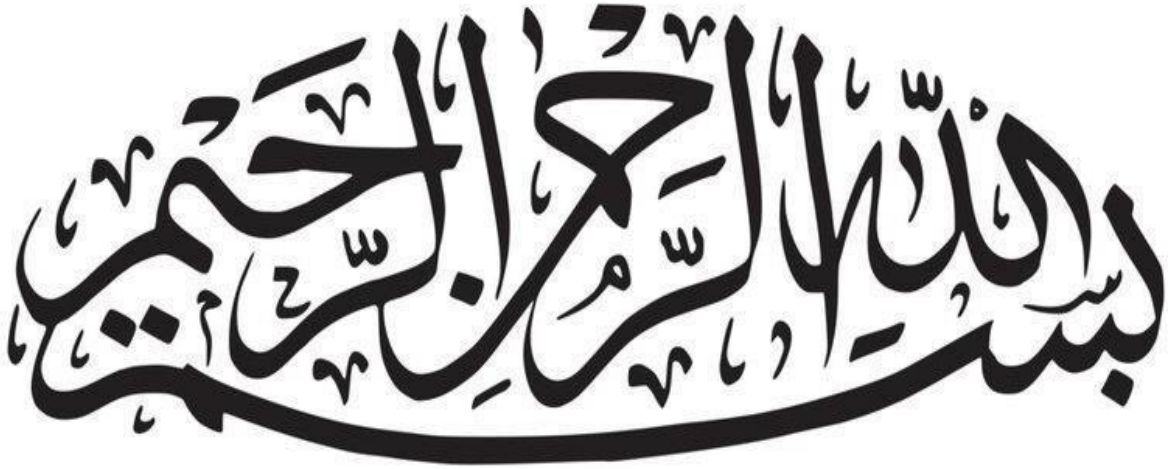
Supervised by
Dr. Attiya Kanwal

Co-Supervised by
Dr. Noreen Akhtar

A Thesis submitted in partial fulfilment of the requirement for the Degree
of
MS in Bioinformatics

Department of Bioinformatics
Faculty of Computing & Information Technology
International Islamic University Islamabad

2025


بسم الله الرحمن الرحيم

DEDICATION

My humble effort, I dedicate to my lovely family and to my respectable supervisor.

Thanks to all of you!

DISSERTATION

A Dissertation submitted to the department of Bioinformatics as a partial accomplishment of the requirements for the awards of the degree of Master of science in Bioinformatics (MSBI)

DECLARATION

We hereby declare that the work presented in this thesis report is our own efforts and that the report is my own composition. No part of the report has been previously presented for any other degree. If any part of this thesis is found to be copied out of any source or Imitation of someone else, I shell stand by the significances

Roman Bibi

ACKNOWLEDGEMENT

I am grateful to Almighty Allah, the most gracious and omnipotent, who has blessed us with the ability to complete our project report. I also express my respect and admiration for the Holy Prophet Hazrat Muhammad (PBUH), who enlightened us with the spirit of one and true Allah and showed us humanity and a genuine way of life.

First and foremost, I am sincerely grateful to my supervisor, **Dr. Attiya Kanwal** for providing me with valuable guidance, support, advice, direction, and encouragement throughout this research journey. Her expertise and mentorship have been a great source of inspiration, helping me grow both academically and professionally. Most importantly, her unwavering support and encouragement in various forms have been a constant source of strength. Without his continuous support, this research project could not have been successfully completed.

I would like to express special thanks to **Dr. Noreen Akhtar** without her contribution and support, completing my research would not have been possible. Her exceptional moral and professional support in completing my research. Furthermore, I am deeply grateful to **Ms. Fatima Iqbal** for her mentorship, and unwavering support during my final thesis. Her guidance was instrumental in bringing this work to completion and I am also grateful to the faculty and staff of the Department of Bioinformatics, their sincere guidance, untiring cooperation, and endless inspiration that enabled us to overcome the entire problems during the course of our studies. Without his continuous support, this study could not have been successfully conducted.

I am extremely grateful to my mother and brother for their love, prayers, care, and sacrifices in supporting my education. They have been there for me both financially and morally, and without their unwavering support, I would not have been able to complete my education. I would also like to express my heartfelt thanks to my sister, and nephews for their love, encouragement, and constant support throughout this journey.

I sincerely thank all my friends, especially Nayab Khalid and Saba Munir, for their kindness and moral support throughout my studies. I am truly grateful for their friendship and the wonderful memories.

Roman Bibi

PROJECT IN BRIEF

The thesis entitled "Leveraging Machine Learning and Molecular Dynamic Simulation to Explore IFRD1 Associated SNPs in Spinocerebellar Ataxia Type 18" submitted by Roman Bibi, Reg No 62-FOC/MSBI/F22 in partial fulfillment of MS Degree in MS Bioinformatics has been completed under my guidance and supervision. I am satisfied with the quality of her work and allow her to submit this thesis for further process to graduate with Master of Science from the Department of Bioinformatics, as per IUI rules and regulation.

Dr. Attyia Kanwal
Thesis Certificate
Assistant Professor
Department of Bioinformatics
Faculty of Computing & Information Technology
International Islamic University, Islamabad

ABSTRACT

Spinocerebellar ataxia type 18 (SCA18) is a neurodegenerative inherited disorder characterized by impaired coordination and balance, requiring a deeper understanding of its genetic basis for effective therapeutic interventions. The dysfunctional cerebellum results in ataxia which can be caused by various factors, such as genetic or hereditary disorders, acquired conditions, toxic or metabolic causes, or structural abnormalities.

In this study, both *in silico* techniques and machine learning approaches were employed to investigate the role of the IFRD1 gene in susceptibility to SCA18 by analyzing the most deleterious SNPs.Insilico analysis was performed using SIFT, PolyPhen-2, SNAP2, PANTHER, SNP&GO, PhD-SNP, MATA-SNP, MuPro, and I-Mutant to evaluate the effects of found SNPs on the protein's function and stability of the proteins. Subsequently, machine learning methods, including Random Forest, Support Vector Machine (SVM), and K-nearest Neighbors (KNN), were utilized to identify significant variations associated with the disease. Integrating both approaches was performed to evaluate the role of IFRD1 in SCA18 susceptibility by analyzing the most deleterious SNPs and their impact on protein structure and stability. Furthermore, molecular dynamics simulations were performed to investigate the effects of these SNPs on protein stability and functionality, providing deeper insights into their potential pathogenic consequences.

The *insilico* analysis of the identified four highly deleterious SNPs based on their predicted impact on protein function. Similarly, the machine learning analysis identified 22 significant SNPs associated with SCA18. The Random Forest model showed the highest predictive performance, with a precision of 62.9%, a sensitivity of 67.8%, and an accuracy of 81%. while SVM and KNN achieved accuracies of 64.8% and 60.4%, respectively improved the reliability of the findings and valuable insights into the genetic factors associated with SCA18. The four identified SNPs rs143002375, rs182917954, rs771235895, and rs1252481308 were found to affect the stability of the IFRD1 protein significantly. Molecular Dynamics (MD) simulations confirmed their substantial impact on the protein's structural integrity, highlighting their potential role in disease development.

The identified functional SNPs offer promising targets for proteomic analysis and therapeutic development, potentially leading to personalized medicine strategies for Spinocerebellar ataxia type 18 (SCA18) patients.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

SCA	Spinocerebellar Ataxia
SCA18	Spinocerebellar Ataxia Type 18 (SCA18)
IFRD1	Interferon-Related Developmental Regulator 1
SNP	Single Nucleotide Polymorphism
nsSNP	Nonsynonymous Single Nucleotide Polymorphism
SNV	Single Nucleotide Variation
RNA	Ribonucleic Acid
DNA	Deoxyribonucleic Acid
MS	Multiple Sclerosis
dbSNP	Database of Single Nucleotide Polymorphisms
SIFT	Sorting Intolerant From Tolerant
SNAP2	Screening for Non-Acceptable Polymorphisms
PhD-SNP	Predictor of Human Deleterious Single Nucleotide Polymorphisms
PANTHER	Protein Analysis Through Evolutionary Relationships
NCBI	National Center for Biotechnology Information
PolyPhen-2	Polymorphism Phenotyping v2
SVM	Support Vector Machine
RF	Random Forest
KNN	K-Nearest Neighbors
ML	Machine Learning
AI	Artificial Intelligence
MD	Molecular Dynamics
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuations
SSE	Secondary Structure Elements
Rg	Radius of Gyration
CDD	Conserved Domain Database
Meta-SNP	Meta-analysis of Single Nucleotide Polymorphisms
EMBL-EBI	European Bioinformatics Institute
RI	Reliability Index

LIST OF FIGURES

LIST OF TABLES

Chapter: 01

Introduction

1. Introduction

This chapter contains comprehensive investigation of spinocerebellar ataxia particular focus on the sub type of spinocerebellar ataxia type 18 on genetic factor that determines the onset of neurodegenerative disease. Subsequently provides an extensive analysis of specific genetic mutation and alter the progression of the neurological diseases, that affect the brain ability to function properly. It emphasizes the role of the IFRD1 gene in the progression of SCA18. These studies highlight the understanding of genetic variation in the IFRD1 gene may lead to alterations in the protein structure, stability and function its encoded protein and, consequently cause the disease.

1.1. Spinocerebellar ataxia (SCA) and SCA18

Spinocerebellar ataxia (SCA) is a genetically heterogeneous group of progressive degenerative diseases characterized by the gradual disintegration of the cerebellum its associated pathways. These disorders are a very heterogeneous group nature marked by the complicated interaction between genotype and phenotype [1]. This degeneration leads to primary motor symptoms, including challenges with speech, balance, and coordination [2]. Spinocerebellar ataxias (SCAs) are inherited in an autosomal dominant. The presence of the disease can be indicated if a person if an individual inherits just one copy of the defective gene from either parent. Several gene have been identified to be associated with different type of spinocerebellar ataxia highlighting the difficulty of the genetic basis of this disease [3]. Globally, of the spinocerebellar ataxia (SCA) prevalence with different incidence rates influenced by the variations in populations and genetic factors. Spinocerebellar ataxia is estimated to affect about 150,000 people in the united states particular time with an estimated 1 to 5 cases per 100,000 individual [4] .This condition is categorized into more than 40 subtype specific genetic abnormalities that affect progression of the disease [5] .The well-characterized subtypes, each associated with unique genetic mutations, include SCA1, SCA2, SCA3, SCA6, SCA17, and SCA18 [6].SCA1 is linked to a CAG repeat expansion in the ATAXIN1 gene which in turn leads to the synthesis of a deleterious protein [7]. Similarly, SCA2 is caused by expansions in the ATAXIN2 gene, through which a harmful protein is produced that interferes with normal brain activity normal brain function. SCA3 is the most common type of spinocerebellar ataxia, of also known as Machado-Joseph disease, is genetically linked to CAG repeat expansions in the ATAXIN3 gene [8].SCA6 and SCA7 SCA6 and SCA7 spinocerebellar

ataxia subtype specific genetic mutation led to a buildup of harmful proteins within the nerve cells that leading to in cellular apoptosis [9]. Other subtypes, such as SCA10 and SCA17, affect normal cerebellar cell function through various genetic pathways, including nucleotide repeat mutations [10]. Furthermore, SCA18 is linked to mutations in the IFRD1 gene, while GRID2 plays a crucial role in the neurodegeneration related to this condition [3].

1.2. Spinocerebellar Ataxia Type 18 (SCA18)

Spinocerebellar ataxia type 18 (SCA18) is a progressive hereditary and genetically heterogeneous neurological disorder primarily affecting the cerebellum, a part of the brain responsible for coordination and balance [11]. The primary symptoms include sensory neuropathy (numbness, tingling, and pain in the hands and feet between ages 13 and 27) and cerebellar ataxia (impairment in coordination, balance, and walking) [12],[13].Spinocerebellar ataxia is a subset of rare inherited cerebellar ataxia disorders. Autosomal-dominant spinocerebellar ataxia typically arising in adulthood are frequently caused by repetitive expansions [14] . Diagnosing SCA18 is challenging due to its rarity and overlapping symptoms with other neurological conditions such as Multiple Sclerosis (MS) [15], Wilson's disease [16] and numerous types of SCAs [17]. Spinocerebellar ataxia 18 (SCA18) typically appears at birth or in early childhood as a genetic disorder. There is a 50% chance that an individual will inherit two mutated copies of the gene, one from each parent [18].At present, the management of the disease highly relies on supportive treatments and therapies to alleviate its symptoms, although a definitive cure has yet to be discovered [19].Although the exact cause of SCA18 is unknown, it is classified as an autosomal dominant genetic disorder meaning that a single defective copy of the gene on an autosome is sufficient to cause the disease. However, the specific causative gene for SCA18 has not yet been conclusively determined [19]. The implication is that the disease can be caused by a mutation of a single gene on an autosomal chromosome (not on a sex chromosome). SCA18 is known to have mutations affecting an estimated one in 20,000 individuals worldwide. However, the specific gene responsible for SCA18 has not yet been identified. Therefore, identifying the genetic basis of SCA18 is crucial for understanding its underlying mechanism and developing targeted therapeutic strategies.

1.3. Symptoms

The symptoms of Spinocerebellar Ataxia Type 18 (SCA18) are likely to be very different from one person to person. The severity differs greatly among individuals in some cases involved, the disease is very slowly progressive form of the disease while others show symptoms much sooner and with more intensity [20]. This particular alteration indicates that in person the SCA18 condition could be manifested differently in each person, and it could show up as very slight problems with balance and coordination to very serious issues with sensory and motor functions. The signs of the disease include gait ataxia, which leads to walking difficulties and unstable mental condition, and sensory neuropathy that leads to numbness, tingling, and poor limb position sense. The symptoms of SCA18 can also be a very different depending on the person affected. Spinocerebellar Ataxia Type 18 (SCA18) symptoms can be very person to person. Spinocerebellar Ataxia Type 18 (SCA18) is depicted in the following illustration. The most frequent symptoms of this disease are disturbances in the senses, and difficulties with motor function, coordination, and balance. The diverse presentation of the disease is indicated by the significant difference in the severity of these symptoms among the affected people **Figure. 1.1** demonstrates the main symptoms.

1.4. Genetic Characterization of SCA18

Spinocerebellar Ataxia Type 18 (SCA18) was first identified by Brkanac *et al.*, in 2002. It was associated with chromosomes 7q22-q32 in a family with ataxia and sensory neuropathy [21] . Spinocerebellar Ataxia Type 18 (SCA18) was identified as a unique subtype of spinocerebellar ataxia in 2013 through investigations into hereditary ataxia disorders. Researchers discovered families that show unique neurological symptoms which are very similar to ataxia and then linked these particular genetic variations with this disease manifestation [22]. Spinocerebellar ataxias, with symptoms such as gait problems, nerve damage, and loss of sensory the researchers studied a specific family showing hereditary ataxia and sensory nerve damage [23].The researcher found the unique genetic alteration in the IFRD1 gene was associated with the symptoms of the family members. This study determined that SCA18 is a unique subtype of SCAs, characterized by the presence of ataxia and peripheral neuropathy [24].

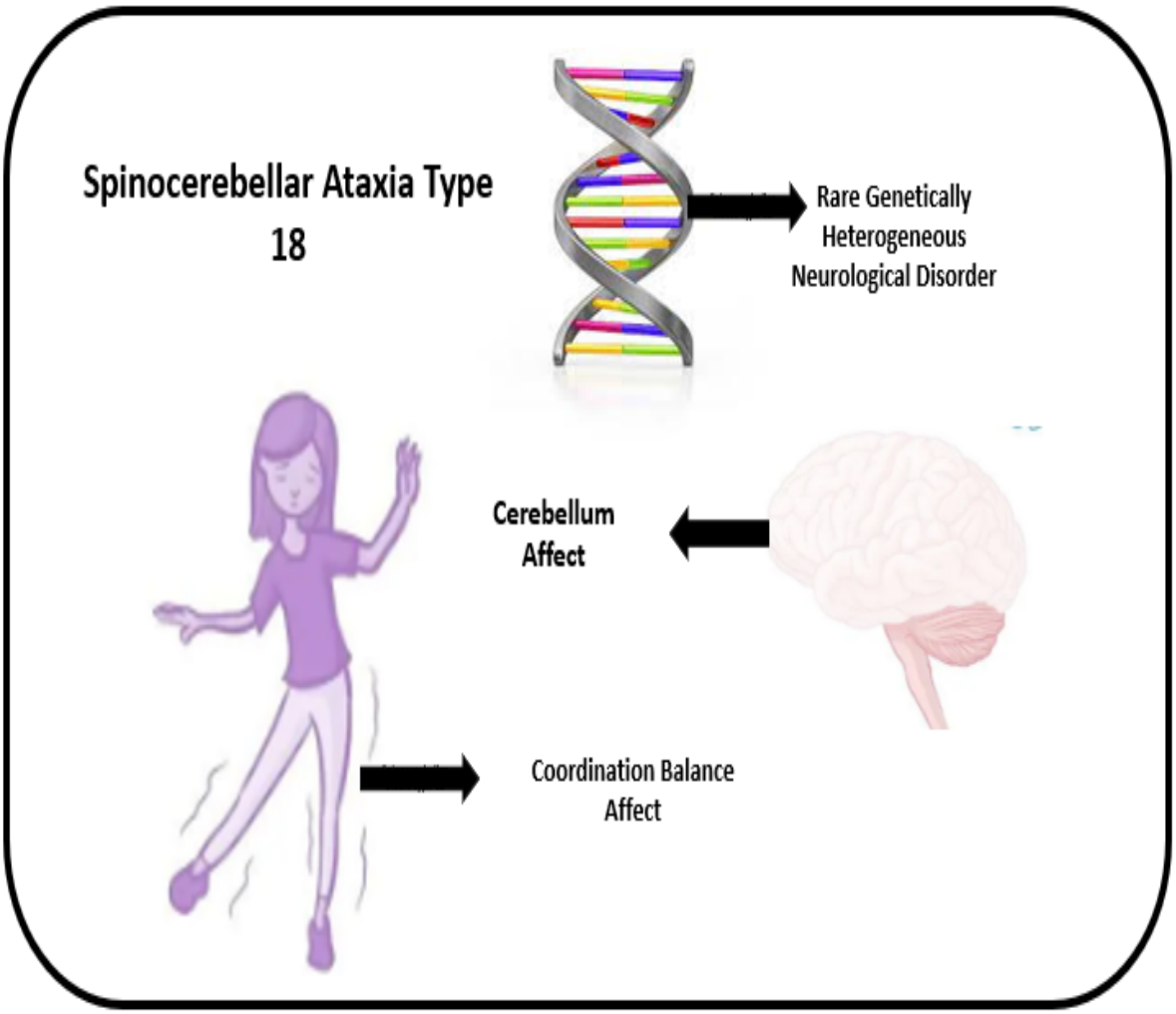


Figure 1.1. The typical symptoms of Spinocerebellar Ataxia Type18

This figure shows the symptoms of Spinocerebellar Ataxia Type 18 (SCA18), highlighting impaired motor coordination, balance difficulties, and progressive neurological abnormalities.

1.5. IFRD1 Gene

Interferon-Related Developmental Regulator 1 (IFRD1) gene is a member to the IFRD family [21].It helps This gene has a significant role in the control of cell division, differentiation (the process where cells acquire specific functions), and reactions to inflammatory signals [25]. With this effect on these functions, the IFRD1 gene is very important in immunological health maintenance and muscular tissue support, particularly in response to physical or immune stress [26]. The human IFRD1 gene is positioned at the 7q31.1 location on the chromosome, has a size of around 18 kb, and has numerous exons and introns. The IFRD1 gene is responsible for the manufacture of a protein implicated in the mechanisms of cell differentiation, muscle growth, and inflammation. The IFRD1 (IFRD1) protein in the human body is composed of up of 451 amino acids and has a molecular weight of roughly 50.2 kDa (kilodaltons) [26].

1.6. Significance of IFRD1 Gene

The IFRD1 gene plays an important role in the regulation of interferon signaling pathways crucial for brain function. These pathways are also involved in the regulation of immune responses, such as cell proliferation and differentiation, and neuronal communication. It regulates the controls the genes expression that are needed for the maturation of these cells into functional forms, and consequently, it affects the overall development and maintenance of the tissue, which directly influence the tissue's overall development and maintenance. IFRD1 is essential for the body's health and functionality, as it provides the development of healthy immune and muscle cells [27]. The gene is predominantly expressed in immune cells, such as lymphocytes and macrophages, while it is also expressed in different tissues. The expression level of the gene is increases in response to interferon and inflammatory stimuli [28]. The IFRD1 protein, which is significant for the development of the embryo, muscle cell renewal and the function of neutrophils, is produced by this gene [29] .Several studies suggested that have indicated that genetic alterations in the IFRD1 gene are associated are linked to various pathological conditions, including cancer, cardiovascular disease, hereditary cystic fibrosis, and neurodegenerative disorders. The significance of IFRD1 in numerous disorders, as well as its potential as a therapeutic target, will provide insight into new treatment techniques SCA18 [25].

1.7. Role in Neuronal Development

The IFRD1 gene plays a significant implicated in a various aspect of neural development, including the development of axons, the formation of synapses, and the neuronal cell maintenance. IFRD1 controlling gene expression to different stimuli including inflammation and oxidative stress which the important for brain health. The role of the IFRD1 in inflammation to give to protective barrier for the neurons against possible damage consequently neurons are easily affected by environmental changes are also impacted in their role in neurodegenerative disease progression [30].

1.8. Variants of the IFRD1 gene

Spinocerebellar Ataxia Type 18 is a genetic disorder associated with the IFRD1 gene it cusses the progressive motor dysfunction characterized by poor coordination, balance, and also with sensory loss. Spinocerebellar ataxias are neurodegenerative illnesses that damage the cerebellum and the circuits that regulate the motions of the body. The mutations found in IFRD1 that are related to SCA18 are considered to have an effect on the pathways that are critical for the functioning of cerebellar neurons, but the exact mechanisms are still unknown [21] . The abnormal regulation of IFRD1 can lead to the susceptibility of cerebellar neurons which together with the motor and sensory symptoms of SCA18. Moreover, IFRD1 has an important role for the protection of neurones from environmental stresses and for the growth of neurons. IFRD1 is associated with neurodegenerative disorders, including SCA18, and can contribute to neurodevelopmental difficulties due to gene mutation or dysregulation [25].

1.9. Global Impact of SCA18

Spinocerebellar Ataxia Type 18 (SCA18) significantly impacts individuals worldwide. results in progressive motor problems and poor coordination, which significantly impact the quality of life, despite its lower prevalence compared to other forms of ataxia [31].Typically, the disorder appears in the early adult years the illness is usually seen in persons of young adult age progressively more care and support are needed as it advances. SCA18 is the unawareness in a lot of countries of the world delays diagnosis and intervention. The current studies are providing new possibilities for better diagnostics and even more new and improved treatments. Especially in the genetic testing and care is limited, the consequences are families and healthcare systems suffering both emotional and financial strains [32]. The global impact of SCA18 in order to addressed by improving healthcare facilities and raising awareness. The

early diagnosis and new treatment possibilities can lead to better outcomes for the affected people [33].

1.10. Variation

Variation concentrates on the distinct non-synonymous single nucleotide polymorphisms (nsSNPs) in the IFRD1 gene these Variation potential impact of the gene function. This study investigates their potential link to Ataxia through the utilization of computational techniques as well as machine learning techniques. Variation is the difference in characteristics between individual members of a species, that can be caused by the genetic or environmental factors. Genetic variations come from mutations, recombination, or gene flow, that supports the process of evolution environmental variance comes from outside sources like climate, nutrition, or lifestyle. Variations play an important role in studies such as genetics, evolutionary biology, and medical research, among others, the comprehension of variations resulted in the identification of the different types of variations. Variation is the primary factor that determines the extent of adaptability, susceptibility to certain diseases, and differences among species. genetic variation two main forms of genetic variation are structural variation and substitution variation [34].

1.10.1. Structural variation

Structural variation refers to major genetic alterations that change the structure of the genome. This includes deletions, duplications, inversions, insertions, and translocations. Such changes can have a strong impact on the function and regulation of genes. Structural variation is an important factor in evolution, increases genetic variation, and is involved caused the onset of diseases like cancer and genetic disorders [35].

1.10.2. Substitution Variation

Substitution variation means that one of the nucleotide base pairs in a DNA sequence is characterized by an alternative base pair or an amino acid in a protein sequence is substituted[36].This change is most often classified as a single-nucleotide substitution, single-nucleotide polymorphism (SNP), or single-nucleotide variation (SNV). The single-nucleotide polymorphisms (SNPs) are divided in to two types: transition SNPs, where one purine (adenine or guanine) is replaced by another purine, or one pyrimidine (cytosine or thymine) by another

pyrimidine, and transversion SNPs, where a purine is replaced by a pyrimidine [37].The genetic variations are very important in genomics research these genetic variations have very significant for the genomic studies, as they provide a basis for genetic diversity, influence the expression of genes, and have associations with different genetic disease and evolutionary mechanism .Categorized as two types of genetic variations SNPs and SNVs the genetic variations are very essential for genomics research [38].

1.10.3. Single Nucleotide Polymorphism (SNP)

Single-nucleotide polymorphisms (SNPs) Single-nucleotide polymorphisms (SNPs) are the major type of variations in the human genome, and these variations consist of a substitution of a nucleotide in either the coding or non-coding of DNA regions [39].Single nucleotide polymorphisms (SNPs) are present in both coding and non-coding areas of the genome, influencing gene expression and protein functionality [40]. In coding regions, SNPs may be classified as synonymous (silent) if they do not modify the amino acid sequence, or as non-synonymous (missense or nonsense) if they affect protein structure or function. Single nucleotide polymorphisms (SNPs) in non-coding areas can affect gene regulation, splicing, or the binding of transcription factors as shown in **Figure 1.2** [41].

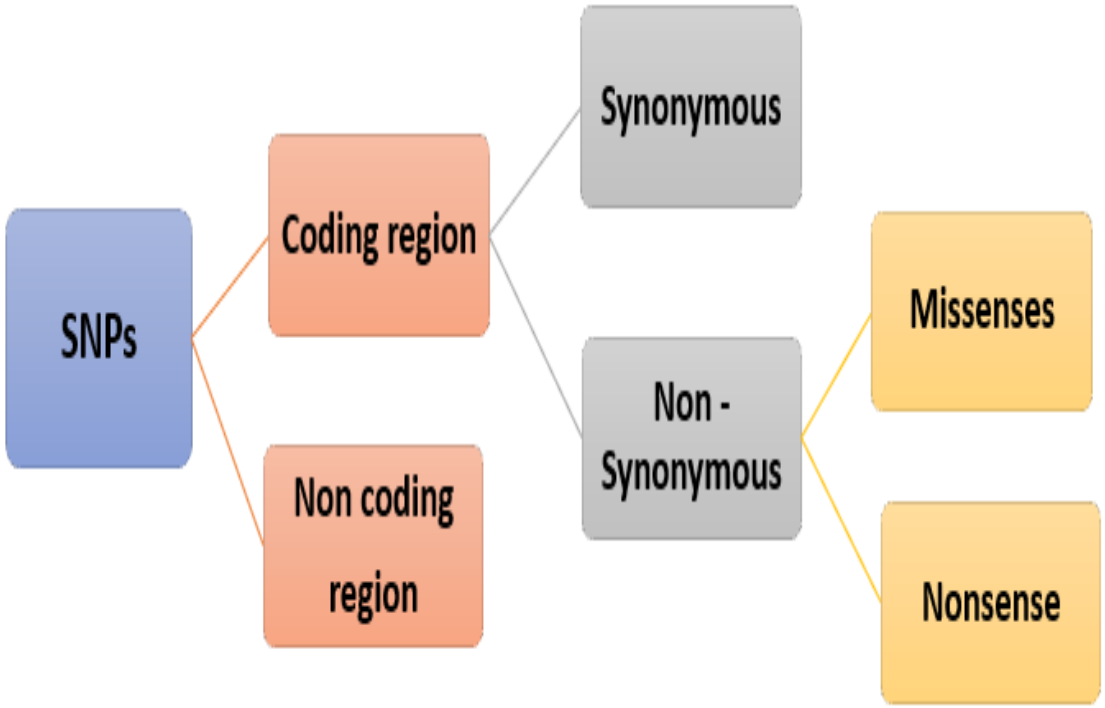


Figure 1.2.A Schematic representation of SNP occurs in both coding and non-coding regions

The figure demonstrates SNPs within coding regions may lead to alterations in protein structure and function. SNPs in non-coding regions can affect gene regulation and expression.

1.11. Role of SNPs in Neurodegenerative Diseases

Single nucleotide polymorphisms (SNPs) are the most prevalent type of genetic variation in humans, affecting only one base location in DNA. SNPs can affect gene activity and increase an individual's susceptibility to diseases, such as neurodegenerative disorders [42]. These variants have a potential to affect gene function and expression, which may lead to malfunction, misfolding, or altered protein synthesis. Many disorders, such as spinocerebellar ataxia, Alzheimer's, and Parkinson's, have been associated with SNPs [43].Approximately 500,000 SNPs have been identified within the coding regions of the human genome [44]. Non-synonymous SNPs (nsSNPs) represent the most common and consequential category of SNPs, often associated with genetic alterations in gene regulation, amino acid composition, transcription factor binding, mRNA stability, cellular and tissue structure, and functional integrity [45]. Furthermore, SNPs significantly contribute to the structural and functional diversity of proteins encoded within the human population [46]. Several SNPs in human genomes have been identified in the non-coding regions of DNA, including the 5′ and 3′ UTRs. Changes in the gene's 5′ UTR can affect transcriptional activity and may have functional significance. Similarly, alterations in the 3′ UTR region can influence gene expression levels by affecting RNA stability or enhancing mRNA translation [47] . Single nucleotide polymorphisms in the IFRD1 gene may substantially affect its function, leading to in neuronal damage. and degeneration in individuals with SCA18. Comprehending these SNPs is essential for elucidating critical elements of SCA18 diseases [48].SNPs are the genetic variation which are responsible for changes in a nucleotide sequence and cause different diseases, including SCA18, breast cancer, Alzheimer's disease, cystic fibrosis, and diabetes mellitus [39].

1.12. Leveraging Machine Learning for Biological Data

The rapid increase of biological data has required the implementation of advanced computational methods for effective analysis and interpretation. Machine learning (ML) emerged as an essential tool for analyzing complex, high-dimensional biological datasets, enabling pattern detection, classification, and predictive modeling [49]. Machine learning applications encompass various biological fields, such as genomics, proteomics, drug development, and illness diagnosis, providing data-driven insights that frequently exceed the efficacy of conventional methods. Machine learning techniques for biological data analysis are primarily categorized into supervised learning and unsupervised learning [50].

1.12.1. Machine Learning Under Supervision

In the supervised machine learning, labeled datasets are used to build predictive models for both classification and regression tasks this feature makes it a suitable technique for gene expression analysis, single nucleotide polymorphism (SNP) classification, and disease prediction [51].

1.12.2. Machine Learning Without Supervision

In the Unsupervised machine learning with unlabeled biological data, performs a very important role in the techniques of clustering, dimensionality reduction, and anomaly detection, which are very crucial for the inference of gene regulatory networks and the prediction of protein structures [52] .

1.13. ML-Based SNP Analysis and Classification in SCA18

A breakthrough tool for evaluating single nucleotide polymorphisms (SNPs), notably in neurodegenerative disorders, is machine learning (ML). The most widespread sort of genetic variation in humans, SNPs have a major influence on an individual's sensitivity to a range of disorders, including Spinocerebellar Ataxia Type 18 (SCA18) with the application of powerful machine learning algorithms [53]. The application of advanced machine learning algorithms to use large genomic datasets to identify SNPs that are related to disease phenotypes, thus improving our knowledge of the genetic factors behind neurodegenerative disorders. Machine learning (ML) is a strong powerful method that is applied to overcome hard problems in different areas and disciplines because of its capability to handle and process high-dimensional datasets. Machine learning has been applied in the analysis of genomic datasets according to numerous Research studies [54]. This methodology instead of comprehensive data mining for the whole area brings to light only some significant biomarkers that are located in areas relatively close to one another To obtain a more comprehensive pool of SNPs for analysis, one can apply dimensionality reduction techniques that are based on random forests (RF) to reduce the size of the dataset before carrying out the cluster analysis RF has been widely utilized in SNP notable characteristics. Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and neural networks are advanced tools that can successfully detect the complex patterns as well as the interactions among the genetic variants [55] .Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and neural networks, make it possible to identify complex patterns and interactions among genetic variants. These

algorithms prioritize single nucleotide polymorphisms (SNPs) based on their ability to predict disease risk, enabling personalized genetic screening and potential therapeutic interventions [51] . Machine learning has the ability to combine different kinds of genomic data, including gene expression profiles and clinical outcomes, thus offering a complete picture of the specific SNPs that affect the disease mechanism [49].The study investigates the SCA18 susceptibility linked with the IFRD1 gene by analyzing harmful SNPs and their impact on protein structure and function using different in-silico bioinformatics tools including machine learning techniques. Machine learning techniques help the identification of SNP data, improving the accuracy of predictions of the SCA18 diseases.

1.14. Problem Statement

Non-synonymous single nucleotide polymorphisms in IFRD1 significantly impact the progression and severity of Spinocerebellar Ataxia Type 18. There is a need to identify pathogenic variants and predict the effects on protein stability of these SNPs and validate them to understand insights into the structural changes in the protein.

1.15. Proposed Solution

This study focuses on the impact of non-synonymous single nucleotide polymorphisms in the IFRD1 gene on protein structure and function, with a particular focus on their potential role in Spinocerebellar Ataxia Type 18 (SCA18). Initially, *in silico* prediction tools, are utilized including SIFT, PolyPhen-2, SNAP2, PANTHER, SNP&GO, and PhD-SNP, to identify pathogenic nsSNPs and evaluate their effects on protein stability and functionality. Subsequently, machine learning algorithms, namely Random Forest, Support Vector Machine (SVM), and k-Nearest Neighbors (KNN), are employed to classify significant variants associated with SCA18, and predictive accuracy. In the final step, molecular dynamics simulations are conducted to validate the structural and stability alterations caused by these mutations, providing comprehensive insights into their pathogenic potential.

1.16. Research Gap

This study highlights IFRD1's role in signaling, regulating cell growth, and stress response. Mutations in IFRD1 are linked to SCA18, as shown in previous studies using in vitro and statistical analyses. The focus is on understanding molecular and structural changes in IFRD1 that drive SCA18 progression through machine learning and insilico analysis to identify potential therapeutic targets.

1.17. Aims Objectives

To investigate the role of IFRD1-associated SNPs in the progression and severity of SCA18, focusing on their identification, functional characterization, and structural impact.

1. To identify single nucleotide polymorphisms (SNPs) within the IFRD1 gene associated with SCA18.
2. To leverage machine learning techniques for predictive modeling and prioritization of pathogenic SNPs.
3. To Investigate associations between prioritized SNPs and disease susceptibility.
4. To analyze the protein stability and functional consequences of identified SNP variants.
5. To validate the predicted structural effects of prioritized SNPs using molecular dynamics simulations.

1.18. Scope of Study

This study analyzes in the role of the single nucleotide polymorphisms in the IFRD1 gene and their linked with spinocerebellar ataxia type 18 (SCA18), rare neurology disorder. The genetic factor involvement in the SCA18 development, we applied the machine learning approaches to investigate the potential impact on the protein function and gene expression, structural stability. To determine the changes in structure and stability of the IFRD1 protein, the selected high-risk single nucleotide polymorphisms (SNPs) underwent molecular dynamics (MD) simulations. The use of these computational methods examines the molecular mechanism of particular SNPs that could potentially impair the activity of the proteins function, provide the foundation for the understanding their contribution to SCA18 progression. The current study focuses on the genetic basis of SNP analysis related to the IFRD1 gene and the SCA18 condition. We applied the molecular simulations and machine learning in this study will allow the identification of possible biomarkers for early diagnosis and therapeutic targets, thus leading to better treatment approaches. Diagnosis and therapeutic targets that could lead to more effective treatment approaches. The Identification of SCA18 and its genetic alterations, enabling the development of precise interventions and ultimately improving clinical for individuals affected by this challenging disorder. The combination of ML and MD in disease-relevant genetic variants' prioritization and validations integration which also contributes to diagnostic and therapeutic strategies for SCA18.

1.19. Impact on Society

The findings of this study have the potential to significantly benefit both the scientific community and society. This research seeks to elucidate the genetic mechanisms behind Spinocerebellar Ataxia Type 18 (SCA18) by focusing on the genetic analysis of IFRD1-associated SNPs. A deeper understanding may lead to improved diagnostic methods and the development of targeted treatments, ultimately benefiting individuals affected by SCA18. This study's mainly positive social effect is its potential to enhance precision medicine. The identification of specific SNPs associated with the condition could facilitate more personalized treatment strategies. Researchers could modify their approaches based on the genetic profiles of individual patients, thereby boosting disease treatment, slowing progression, and improving the quality of life for individuals with SCA18. Additionally, this research may facilitate the advancement of novel therapy alternatives for SCA18, addressing a significant deficiency in current therapy. By focusing on the genetic factors that contribute to neurodegenerative diseases, targeted treatments could provide more viable and efficient solutions, as most current therapies offer only limited aid. This may contribute to an improvement of the healthcare burden associated with the disease and the enhancement of overall patient treatment.

Chapter: 02

Literature Review

Literature Review

In this literature review chapter, we will cover an extensive examination of the single-nucleotide polymorphisms (SNPs) that have been linked with Spinocerebellar Ataxia Type 18 (SCA18) disease, with a particular focus on the alterations of the IFRD1 gene. This study will explore the associations of these SNPs may be impact on the development of the SCA18 disease as well as the functional consequences of these variation. Moreover, the current research focused on the analysis of various methodologies applied in recent research studies which such as machine learning, computational method and molecular dynamic simulation to assess the impact of the genetic variation. This chapter highlights substantial changes in this area along with important gaps that need to be filled further investigation. This literature review Will enhance the knowledge of the genetic factors involved in SCA18 and possibly their impact on disease control and treatment.

2.1. Spinocerebellar Ataxia

Sullivan and colleagues discussed 2019 a group of hereditary neurodegenerative disorders were characterized by gradually progressive cerebellar ataxia. Such disorders are majorly due to CAG repeat expansions or similar genetic changes. The development of next-generation sequencing (NGS) technologies has improved the diagnostic capacity, thereby allowing the discovery of new genes and mechanisms, for instance, the harmful RNA gain-of-function, the mitochondrial malfunction, and the poor autophagy. In this research on genetic modifiers that influence disease onset and progression is active, thus, the role of personalized treatment in SCA management is emphasized [5] .

2.2. Cause of Mutations (SCAs)

The article published by Müller *et al.,* in 2021 discusses the resulting difficulties in coordination and movement. Müller 2021 presents a thorough study of the genetic aspect of SCAs and points out that these disorders are often caused by mutations in certain genes, including expansions of trinucleotide repeats or point mutations that impair significant cellular activities. Spinocerebellar Ataxia Type 18 (SCA18), a rare autosomal dominant disorder, is one of the various SCAs that affect the 7q22-q23 chromosomal region because of the mutations. The findings broaden our comprehension of the genetic intricacy involving SCAs

and also imply new potentials for gene-targeted therapy in the case of these diseases[6].

2.3. Case Study 1 for SCA18

The articles publish by Hetzel *et al.,* in 2020 the main focus of the research was on SCAR18, an extremely rare neurological disease that is caused by bi-allelic mutations in the GRID2 gene. GRID2 is the gene encode delta-2 ionotropic glutamate receptor. The research identifed particular homozygous nonsense mutation variant c.568C>T; p. Gln190 that was linked to the with the most sever clinical symptoms in the patient, such as early-onset cerebellar ataxia, significant developmental delay, neurological difficulties, and cerebellar hypoplasia. SCAR18 also highlighted the possibility of genotype-phenotype correlation through the GRID2 mutations which was a significant contribution better understanding of this rare condition. The results of this research could play major role in the development of more efficient diagnostics, therapies, and management strategies for patient with this rare neurological disorder [56].

2.4. Case Study 1 for SCA18 2

The paper by Panda *et al.,* 2022 discusses an investigation that analyzed an SCA18 leads to the gradual development of ataxia, dysarthria, and cerebellar dysfunction. The authors explored the genetic factors of SCA18, mainly the particularly the role of GRID2 mutations and their effect on neural activity. The author examined the thorough genetic sequencing, clinical evaluations, and the neurological diagnosis to investigate the phenotype-genotype relationships in SCA18. In order to clinical feature and symptoms were associated with genetic variation and the strategy consisted of analyzing existing case reports and studies. This paper highlights the significance of genetic testing in diagnosis but also provides the valuable information about with SCA18 occurrence associated with GRID2 mutations. The findings emphasize not only the need for early identification and treatment of the disease, as well as for a better understanding of the clinical spectrum and genetic basis disease. Further investigation is essential to explore genetic treatments and improved therapeutic treatment of SCA18 [57].

2.5. SCA Spectrum

Manto *et al.,* 2005 conducted a study that investigated spinocerebellar ataxias (SCAs)

in with a similar prevalence of approximately 1-4 per 100,000 people. Spinocerebellar ataxias (SCAs) are a disorder that affects the cerebellum and causing various problems in the body functions such as coordination, speech, eye movement identification of SCAs is problematic because to overlapping features across genetic groups. The SCA group of spinocerebellar ataxias will provide very important clues into the diagnosis of the diseases. Among them are the slowing of nerve impulses in SCA2, ophthalmologic in SCA1, SCA2, and SCA3, pigmented retinopathy in SCA7, and spasticity in SCA3. The alterations in the genes responsible for spinocerebellar ataxias (SCAs) re primarily caused by tetra nucleotide repeat expansions, specifically CAG repeats, which are responsible for the different forms of ataxia such as SCA1, SCA2, SCA3, SCA6, SCA7, SCA17, and SCA18.With genetic testing finding causative genes in 60–75% of cases, molecular understanding of SCAs is quickly increasing, offering hope for future viable treatment methods [58].

2.6. IFRD1: Candidate Gene for SMNA

In the investigation done by Brkanac, Z., *et al.*, 2009 localized to chromosome 7q22-q23. According to the research, this genetic disorder is caused by mutations in the IFRD1 gene. Brkanac and colleagues revealed that mutations in the IFRD1 gene may dramatically modify the etiology of sensory/motor neuropathy with ataxia (SMNA), marked by neurological symptoms including muscle weakness, poor movement coordination, and sensory abnormalities. Additionally, their findings increase our understanding of the genetic components of SMNA and will contribute in the development of targeted therapies and interventions in the future[26].

2.7. SMNA Linked to Chromosome 7q22-q23

The paper by Brkanac *et al.*, 2002 discusses an investigation that analyzed an autosomal dominant neurological disorder associated with sensory and motor neuropathy with ataxia (SMNA), which is characterized by the presence of sensory and motor neuropathy and ataxia. The analysis indicates a major correlation between the disorder and the chromosomal area defined as 7q22-q23, which is considered to be the genetic basis of the disease. The genetic study of the disease also involves the investigation of affected families in which a specific chromosomal position that may contain the genes linked with the disease. Further study into

the pathogenesis of sensory/motor neuropathy with ataxia (SMNA) and new therapies is made feasible by their findings, which expand our awareness of the genetic mechanisms underlying SMNA [21].

2.8. IFRD1 Variant in Hereditary Spastic Paraplegia Ataxia

Lin *et al.,* 2018 study provided to uncover a particular mutation in the IFRD1 gene associated to the family's clinical symptoms, the research employed a thorough genomic method. A Han Chinese family with autosomal dominant hereditary spastic paraplegia, peripheral neuropathy, and ataxia was revealed to have a missense mutation in the IFRD1 gene (c.514A>G, p.I172V). The research has indicated a possible correlation between the IFRD1 gene and specific neurological disorders. These findings the role of the IFRD1 pathogenesis of hereditary spastic paraplegia and imply that gene alterations in the gene lead to a variety of neurological disorders. The findings underscore the relevance of genetic research in elucidating the complexity of inherited neurodegenerative diseases and enhance knowledge of the genetic components connected to these disorders [25].

2.9. IFRD1 Polymorphisms in Cystic Fibrosis

The research paper Baldan *et al.,* published in 2015, mainly focus to investigate the relationship between specific genetic variations of the IFRD1 gene with nasal polyps as well as the diagnosis of cystic fibrosis (CF).In the study of 143 patients with cystic fibrosis, in which 40 patient of the nasal polyps, through an analysis three harmful SNPs (rs7817, rs3807213, and rs6968084) identify in the IFRD1 gene .Thus, the alterations in the IFRD1 gene may impact on the inflammatory processes that lead to the development of the polyposis and its association with cystic fibrosis. This research emphasizes the need for particular genetic testing confirm genetic susceptibility of cystic fibrosis patients to respiratory problems is one of the main contributions of this research paper [59].

2.10. IFRD1 polymorphisms in Gastric Cancer

In the 2014 study conducted by *Xu et al.,* The authors studied several single nucleotide polymorphisms in the IFRD1 gene—rs7818, rs3807213, and rs6968084—among 53 patients with cancer of the gastric and 50 healthy individuals. The results of their research indicated

that the rs3807213 C allele and the rs3807213 C/C genotype were considerably more common in gastric cancer patients than in the control group, implying the link among these particular IFRD1 snps and an elevated risk of gastric cancer development. However, no substantial associations have been identified between the remaining two SNPs (rs7818 and rs6968084) and the susceptibility to stomach cancer. This work emphasizes the potential significance of IFRD1 genetic variants in the etiology of cancer of the stomach and highlights the necessity for further studies to investigate IFRD1 as a potential diagnostic for evaluating gastric cancer risk [60].

2.11. IFRD1 Polymorphisms in Cystic Fibrosis Neutrophil

Gu *et al.,* 2010 focused in their study particularly on the effect of neutrophils' activity. The research confirmed that these genetic variations have an impact on neutrophil reactions, which in turn led to inflammation and increased cytokine production in cystic fibrosis patients. The immune response that is not well-regulated is a major cause of the lung damage that occurs in cystic fibrosis. The authors suggested that the IFRD1 gene variations may serve as potential markers for predicting the severity of inflammation in cystic fibrosis and could steer personalized treatment approaches that concentrate on the regulation of the immune system. The major aspect of this study is that it focuses on immune regulation. The review brings forth the crucial role of IFRD immunological dysregulation and its potential for therapeutic intervention in cystic fibrosis, namely through treatments designed to enhance neutrophil activity and mitigate chronic inflammation [61].

2.12. GABR SNPs in Neurodevelopment

Manaz *et al*., 2023, in their research which focuses on analyzing to examine the potential consequences of missense single nucleotide polymorphisms (SNPs) in the GABRA1, GABRB1, and GABRB3 genes. These three genes encode for the different types' subunit type respecter GABA_A receptor. The GABA_A receptor that play an important role in the inhibition of neurotransmission within the central nervous system. The author uses the different Bioinformatics tool Such as SIFT, Polyphen2, PROVEAN, PANTHER, MutPred2, Predict SNP, PhdSNP, PMUT, and SNPs & GO these tools predict the possible effects of these mutations on the protein structure and function. These analyses identify different SNPs that

may have effect on the protein function in the play important role for the neurodevelopmental disorder progression. The studies emphasize the value of *in silico* approaches in finding the molecular foundations of numerous disorders and suggests promising candidates for future experimental exploration [62].

2.13. *in silico* Prediction of RASSF5 SNP Effects

Hossain *et al*., 2020, in their research which is critical for tumor suppressor regulation. In the author's study Employing in silico techniques, they found that these mutations may likely change the structure and function of the RASSF5 protein, thereby altering biological processes such as apoptosis and the cell cycle. The authors employ various bioinformatics methods and the application of molecular dynamics simulations to identify the harmful SNPs that influence the protein structure and cellular molecule function. To the association of RASSF5 gene variation with cancer susceptibility he significance of computational methods in forecasting the functional consequences of genetic alterations. In this paper highlighted the importance of these method elucidation of the molecular pathways of diseases and for the eventual development of possible interventions [63].

2.14. Computational Analysis of Deleterious nsSNPs in p14ARF

Ahmad *et al.*, 2024 in their research which focuses on analyzing to examine the potential consequences of missense single nucleotide polymorphisms (SNPs) n the p14ARF (CDKN2A) gene. The authors employ various bioinformatics methods and the application of molecular dynamics simulations to identify the harmful SNPs that influence the protein structure on the p14ARF protein. These analyses identify different SNPs that may have effect on the protein function in the play important role for influencing cellular processes and boosting cancer risk. The findings highlighted the importance of computational strategies, in particular molecular dynamics simulations, in predict the effects of genetic variations, provided deep insight into the molecular mechanisms that control genetic abnormalities and diseases like cancer [64].

2.15. *In silico* Analysis of nsSNPs in CTLA4

Irfan *et al.*, 2023, in their research which focuses on analyzing to examine the potential

consequences of missense single nucleotide polymorphisms (SNPs) in the CTLA4 gene which is a significant regulator of the immune system and has been linked with autoimmune diseases. The author uses the different Bioinformatics tool Such as SIFT, Polyphen2, PROVEAN, PANTHER, MutPred2, Predict SNP, PhdSNP, PMUT, and SNPs & GO these tools predict the possible effects of these mutations on the protein structure and function. These analyses identify different SNPs that may have effect on the protein function in the play important role for the leading to immune imbalance. The author uses different bioinformatics tools to analyze the impact of genetic variation on functional changes so that potential genetic risk factors for autoimmune diseases. The research highlights the role of such methods in understanding the molecular mechanisms of diseases and therefore, in selecting the appropriate drug treatment strategies[65].

2.16. *In silico* Analysis of nsSNPs in IL-10

The study by Das *et al.*, 2022 highlighted the importance of genetics in the immune system control mechanism. The author identifies these genetic modifications can have a strong impact on the protein's protein structure and function regulating the inflammatory responses. The alterations could possibly increase the risk of developing different conditions, including autoimmune diseases or inflammatory disorders. the author has highlighted the relevance of these nsSNPs in health and disease by employing various bioinformatics tools for the functional analysis of the nsSNPs they identified the impact of these mutations in health and disease. The study understanding genetic differences regulating immune responses it will provide new therapeutic approaches for the diseases associated with immune modulation [66].

2.17. Machine Learning for SNP-Based Asthma Prediction

The research paper by Gaudillo, J., and colleagues 2019 on the use of machine learning techniques for the of forecast risk of asthma associated with specific single nucleotide polymorphisms has a new path for personalized therapy. They first used Random Forest (RF) and Recursive Feature Elimination (RFE) for the purpose of identifying important SNPs that are related to asthma risk. After that k-Nearest Neighbor (kNN) and Support Vector Machine (SVM) techniques were used for the classification and of asthmatic differentiation of asthmatic and non-asthmatic persons, respectively. The study indicates that the integration of machine

learning forecasts and genetic data could lead to more accurate disease prognosis, thereby opening a new path for individualized therapy. The study also indicates that these techniques could lead to improved early detection, and diagnosis, and management of asthma, thus increasing the effectiveness of precision medicine [53].

2.18. Machine Learning for SNP-Based Precision Medicine

In 2019, Daniel Sik Wai Ho conducted research that was the main focus was the prediction of the single nucleotide polymorphisms (SNPs) consequences in the help of in personalized medicine through the application of machine learning techniques. The researchers first conducted SNP selection used ensemble technique and then applied gradient boosting trees (XGBoost) to tackle the problem of non-linearities and interaction effects. This technique was shown to be precise over nine complex phenotypes in a multi-ancestry class, whereby its potential to enhance disease risk prediction models was demonstrated. The paper accentuates the necessity of integrating functional annotations and population-specific SNP datasets to enhance model accuracy. The authors emphasize that the combination of genomic data with powerful machine learning algorithms can produce fresh insights into personalized medicine through the discovery of genetic variants associated with complex traits. Moreover, the they highlight the difficulties of the interpretability of machine learning models and recommend improvement as interpretability issues and suggest that more work is needed to make these models clinically applicable. The studies give an indication of the ability of machine learning applied to genomic data to facilitate personalized medicine[49].

2.19. ML-Based SNP Analysis for Disease Susceptibility

Roxas-Villanueva *et al.*, 2022 conducted a machine learning based study to identify the disease-associated loci and subsequently, the nature of the patients' susceptibility to the given disorder. The researchers employed Random Forest (RF) classifiers along with clustering techniques in a combined approach, thereby allowing the integration of complex genetic data so as to improve the predictions of the disease risk. The experiment was based on the sevoflurane data of hepatitis B virus surface antigen (HBsAg), which led to the revealing of new patterns regarding the susceptibility of individuals based on their genotype. The authors highlight that the comprehension of the complex disorders and suggest that their approach

could be a significant tool in the precision medicine. Machine learning applied to the combination of multi-layered omics data is considered by them to provide the best opportunity for detection of essential biomarkers for early diagnosis and treatment[50].

2.20. GWAS Asthma Prediction

M.Xu *et al.*, 2011 performed a Genome-Wide Association Study (GWAS) to forecast severe asthma exacerbations in children utilizing Random Forest classifiers. Their research intended to uncover genetic variations linked to asthma severity by utilizing machine learning methods to enhance prediction precision. The authors demonstrated the efficacy of Random Forest models in identifying critical genetic markers associated with asthma exacerbations through the analysis of extensive genomic data. Their findings highlighted the potential of combining GWAS with machine learning to reveal genetic predispositions to complicated disorders. This research advanced personalized therapy by facilitating improved risk assessment and focused therapeutic methods for pediatric asthma patients [67].

2.21. AI & ML in Precision Medicine

In the paper by Joshi *et al.*, 2021, entitled “Integration of Artificial Intelligence, Big Data and Machine Learning in Precision Medicine and Drug Development,”, the authors investigate the role of AI, big data, and ML in the new era of precision medicine and drug development. According to the authors, modern technologies are the ones that make it possible to analyze large amounts of biological data, which helps in discovering new candidates for drugs and tailoring treatments based on the patient's genetic profile. The applications of AI and ML in such a way allow the researchers to not only to predict the effectiveness of the treatment but also to speed up the process of drug development and to create more effective treatment plans for the patients. The study also mentions the issues of data quality, integration challenges, and the need for robust computational models as some of the main challenges. The research gives the hope for the enhancement of the use of AI and ML mainly in personal health care and personalized medicine [68].

2.22. SVM-Based Protein Stability Prediction

In 2006 Cheng and colleagues discussed a group of a study that focused on the

development of a computational model to that could distinguish between the effects of stabilizing and destabilizing mutations using various protein and structural feature. The research enhances the precision of mutation effect assessment, the predictive model based on SVM that also such as a physicochemical properties and evolutionary data. The model was verified with the help of experimental data, and it was found that the SVM method was able to make predictions of the protein stability changes with a greater reliability than the classical statistical methods. This study finding highlighting the importance of machine learning approach in understanding gene mutation and their possible impact on the protein function structure and progression of the disease mechanism and drug discovery. This paper provides a fundamental tool design more challenge to investigate further in computational protein and harmful mutation prediction [69]**.**

Chapter Overview

The current investigation into IFRD1 gene role in Spinocerebellar Ataxia Type 18 (SCA18) diseases employs a variety of techniques,including SNP Analysis in vitro studies. These approaches have uncovered several important findings regarding IFRD1 involvement in theseconditions. Firstly, IFRD1 is strongly associated with increased Spinocerebellar Ataxia Type 18 (SCA18). Research on the IFRD1 gene and its associated SNPs demonstrates potential for therapeutic applications. However, further validation is essential to confirm their functional significance and clinical efficacy.

In particular, studies focusing on the IFRD1 gene have highlighted its significant role in regulating the aggressive phenotype of RAFLS. Additionally, research has demonstrated that disruptions in IFRD1 can contribute to the pathogenesis of Spinocerebellar Ataxia Type 18 (SCA18), linking it to neurodegenerative processes. This suggests that targeting IFRD1 could potentially be a therapeutic strategy in neurodegenerative diseases such as SCA18. The role of IFRD1 in regulating disease phenotypes across these conditions opens new avenues for research into its therapeutic potential in d neurodegeneration.

Current research efforts mainly concentrate on SNP-based analysis, machine learning methodologies, and bioinformatics tools to forecast genetic variants associated with diseases and their functional implications. This biomarker could enhance diagnostic accuracy and help treatment strategies. Studies on IFRD1 and its association with SCA18 highlight the potential influence of SNP variants on disease progression.

Table 2.1. Comparative Literature Outcomes for SCA18

Authors/Year	Title	Methodology	Outcomes	Limitations
(*Müller et al.,* 2021)	Spinocerebellar Ataxias (SCAs) Caused by Common Mutations	A comprehensive review of existing knowledge on genetic sequencing identify mutation SCA and sub type sca18 its role in diseases development	SCA and SCA18 are associated with novel pathogenic mutations, emphasizing its contribution to neurodegeneration.	IFRD1 needs to be targeted for the identification of better therapeutic agent againstSCA18
(*Hetzelt et al,*2020)	A Case of Severe Autosomal Recessive Spinocerebellar Ataxia Type 18 with a Novel Nonsense Variant in GRID2	Clinical studies to identify the impact of a novel nonsense variant linked to SCA18	Found novel nonsense variant in *GRID2* that are responsible of SCA18	Molecular dynamic simulation techniques need to explored the IFRD1 variants better Therapeutic Gent againstSCA18
(Brkanac et al., 2009)	IFRD1 Is a Candidate Gene for SMNA on Chromosome 7q22-q23	In vitro experiments using human cell lines were performed to evaluate the functional impact of identified genetic variants. In vitro experiments on human analysis disease prediction	It was found that a nonsynonymous variant in the human interferon-related developmental regulator gene 1 (IFRD1) has been identified as a potential disease-causing candidate.	Insilico analysis of the IFRD1 mutations

(Brkanac, Z. *et al.*,2002)	Autosomal Dominant Sensory/Motor Neuropathy with Ataxia (SMNA): Linkage to Chromosome 7q22-q32	In vivo and In vitro experiments to study the Involved the collection of DNA samples from both affected and unaffected family members across numerous generations.	The identification of Sensory/Motor Neuropathy with Ataxia (SMNA) disorder	Molecular modeling for the IFRD1 gene.
(Xu et al., 2014)	IFRD1 polymorphisms and gastric cancer risk in a Chinese population	Using wet lab techniques to study theIFRD1 gene mutation in gastric cancer patients	It was found that a IFRD1 gene mutations associated with gastric cancer risk.	Identify a biological target for the IFRD1 gene sca18
(Gu et al., 2010)	IFRD1 polymorphisms in cystic fibrosis with potential link to altered neutrophil function.	In vitro human studies Statistical models were utilized to associate genetic variants with disease severity.	IFRD1 polymorphisms found were significantly associated with variation in neutrophil effector function.	Computational analysis of the IFRD1gene
(Daniel *et al., 2019)*	Machine Learning SNP Based Prediction for Precision Medicine	The study utilized supervised machine learning models trained on single nucleotide polymorphism (SNP) data to forecast particular disease risks.	The machine learning models exhibited enhanced accuracy in predicting disease risk relative to conventional polygenic risk scoring techniques.	IFRD1 genetic data into predictive models Their potential in personalized assessment to advance precision medicine SCA18.

(Gaudillo, J., *et al.,* 2019)	Machine Learning Approach to Single Nucleotide Polymorphism-Based Asthma Prediction	The study used Machine learning techniques to examine SNP data and forecast asthma susceptibility.	The models precisely identified SNPs linked to asthma risk, Highlighting their potential as biomarkers for disease susceptibility	The need to the IFRD1 Feature selection and classification method enhanced predictive performance for therapeutic innervation for the SCA18 disuses.

:

Chapter:03

Materials and Methods

3.Materials and Methods

This study presents the methodological framework that was used to investigate the impact of single nucleotide polymorphisms (SNPs) associated with the IFRD1 gene IFRD1 gene in Spinocerebellar Ataxia Type 18 (SCA18).The methodology consists of a step-by-step process starting from SNP data collection and then using machine learning to predict the functional impact of the variants .Afterwards, the cross-validation techniques are employed to confirm the reliability and accuracy of the predictive models. Structural effects of SNPs on the IFRD1 protein were identified to molecular dynamics simulations.

3.1. Collection of SNPs Dataset

The SNP dataset was obtained from various genetic sources that are publicly available, thus providing a comprehensive foundation for the assessment of the genetic factors associated with the disease.

3.1.1. dbSNP

The dbSNP database was used to obtain form the SNP-related information such as SNP ID, protein accession number, location, and altered residue. DbSNP (Database of Single Nucleotide Polymorphisms) is the most extensive public repository that that collects data on genetic variants and specifically, small-scale genetic modifications such as insertions, deletions, and SNPs data. It's provided the genetic information of SNPs ID location, alleles associated with genes, frequency which helps to understand genetic differences and their relation to diseases, evolution, and diversity among humans [70]. dbSNP database is a comprehensive, and freely available database for single nucleotide polymorphisms and genetic variation. The identification and analysis of genetic variants genetic diversity, association with diseases, and the functional impact of these differences in various populations [71].

3.1.2. Ensembl

The Ensembl database from the European Bioinformatics Institute (EMBL-EBI) is a publicly freely available database used for the genomic genomics data. The database provides information about the gene sequence, genes, protein-coding sequences, protein structures, regulatory elements, and variations in genes form different species across a large number of annotated genomes. Ensemble collects data from different sources which help understand gene roles, evolution and diseases. The commonly used methods for such analyses include variant

analysis, comparative genomics, and functional genomics [72]. Ensembl has given access to large collection genomes data, variants, and their biological significances. Ensembl provides resources for genome-wide association studies (GWAS) of genetic analysis [73].

3.1.3. Polysearch

Polysearch friendly and open-source database that is used for the retrieval of genomic data. The database gives the extensive analysis of the impact of variant and function of gene. Polysearch was utilized for identifying the possible consequences of variants and mutations on the disease development. the database gives access to genetic information, functional annotations, and evaluate the possible impact of the genetic variations and phenotypic effects phenotype consequences [74].

3.1.4. DisGenet

The DisGenet open-source database was employed for retrieve data on Gene-Disease relationships. The processes of retrieval and filtering data with different levels of curation and the corresponding confidence scores to ensure the results is accurate. In order to the biological implications of these genes to be understood, the refined data was further analyzed using functional annotation techniques. The single nucleotide polymorphisms (SNPs) were classified into different types based on their location in the genome, such as the in-frame indel, in-farm insertion, initiation codon variant, frameshift, in-frame deletion, intronic, missense, non-coding transcript variant, and synonymous mutations [75].

3.1.4. NCBI

The NCBI (National Center for Biotechnology Information) open-source database give the wide range of biological and genetic data. The tool was utilized for the purpose of data collecting, analyzing, and interpreting genetic data obtained from different sources such as NCBI Gene, dbSNP, PubMed, OMIM, and ClinVar. The significance of these databases for the identification of mainly related literature, mutations, and genetic variations linked to different diseases. In order to evaluate the importance of genetic variations and their biological influence collected genetic data through processed, filtered, and analyzed data. The IFRD1 protein sequence in FASTA format was obtained from the NCBI (National Center for Biotechnology Information) database [76].

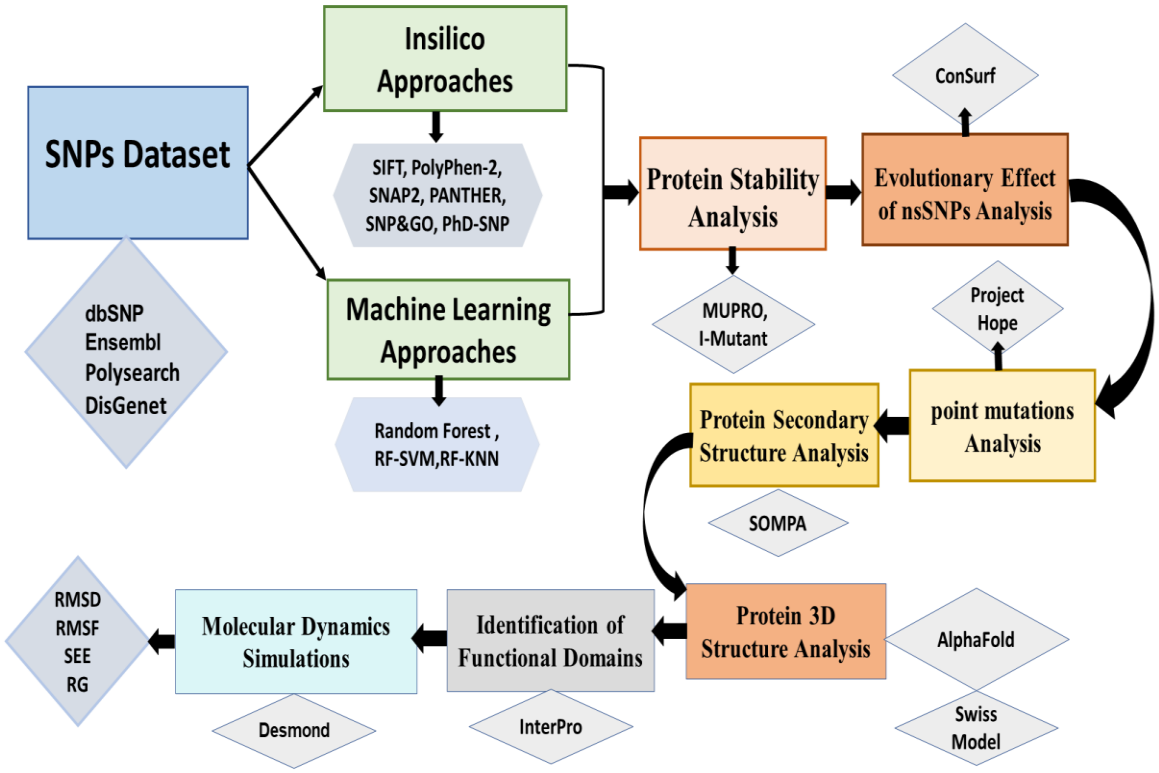


Figure 3.1. Schematic diagram Illustrating the identification of SNPs in the IFRD1 Gene

3.2. *Insilico* Approaches

SNP analysis is the process of identifying single nucleotide polymorphisms and interpreting them which help to reveal the genetic variations related to diseases and traits. By employing the various computational bioinformatics tools for SNP analysis which in help the prediction of harmful mutations and the determination of possible impact on biological function. We used six different bioinformatics tools: SIFT, PolyPhen2, SNPs&GO, PANTHER, Meta-SNP, and SNAP2 to assess the structural and functional impact of non-synonymous single nucleotide polymorphisms (nsSNPs) on the IFRD1 protein. The outputs from these six tools are identify in the high probability causing a significant change in protein function.

3.2.1. SIFT

The SIFT (Sorting Intolerant from Tolerant) tool is utilized to predict the functional impact of amino acid or nucleotide substitutions based on evolutionary conservation across related protein sequences[77].SIFT, a tool that predicts the effect of amino acid substitutions on protein function, was applied to classify the mutations in a dataset of non-synonymous SNPs (nsSNPs) obtained from the NCBI dbSNP database as tolerated or deleterious [78].The scoring of the algorithm assigns based on the scale 0.0 (harmful) to 1.0 (accepted) with the variations that get the score of 0.0 to 0.05 being categorized as functionally detrimental. the Higher scores indicate a reduced likelihood of adverse functional consequences [79].

3.2.2. PolyPhen-2

PolyPhen-2 (Polymorphism Phenotyping) bioinformatics tool that is commonly used to predict the impact of non-synonymous single nucleotide polymorphisms (nsSNPs) on the human protein structure function [80]. PolyPhen-2 predicts the potential effect of single nucleotide polymorphisms (SNPs) on human genetic variation by using both structural and sequence-based features. It also considers several aspects such as amino acid physicochemical properties, their location in the protein, and their conservation across species [81]. The pathogenicity of missense variations is considerably revealed by the prediction of the probability that a specific amino acid alteration is either benign or harmful to protein function. The PolyPhen-2 prediction score can vary from 0 to 1, indicating the harmful mutations respectively[82].

3.2.3 SNPs&GO

SNPs&GO is a freely available Web server that predict with specific the effect of single point mutations, which are mainly single nucleotide polymorphisms (SNPs), as related to diseases. The SNPs&GO is a support vector machine-based (SVM) web tool used for SNP pathogenicity prediction. It utilizes gene ontology and protein domain annotations to gather information and consequently, it proficiently classifies SNPs into two categories, disease-causing or benign. Prediction score between 0 and 1[83].

3.2.4 PANTHER

PANTHER (Protein Analysis Through Evolutionary Relationships) is a broad bioinformatics resource that provides valuable insights into proteins classification, gene function, and pathway analysis[84]. It classifies proteins into families and subfamilies according to on their evolutionary associations. PANTHER determines the probability that a mutation at a particular nucleotide or amino acid residue would impact protein function, showing that mutations in extremely prevalent regions are typically deleterious [85]. To predict single nucleotide polymorphisms (SNPs) by employing evolutionary insights and functional associations. The score range gives 0.1 between 0.5. The ratings surpassing 0.5 imply a possible detrimental impact[86].

3.2.5 Meta-SNP

Meta-SNP (Meta-analysis of Single Nucleotide Polymorphism) computer software, based on a meta-analysis technique, predicts the influence of SNPs or genetic variants on proteins. Meta- SNP is vital for analyzing the genetic potential dangers related with certain SNPs [87].It gives key insights into these variations in genes may alter biological activities as well as contribute to disease processes, guaranteeing that these relationships are verified. The meta-score generated by Meta-SNP swings between 0 and 1. The scoring approach is based on a user-defined threshold of 0.5 to pick SNPs with a high chance of functional impacts, whilst a threshold of 0.1 is used to identify potentially intriguing SNPs that warrant additional experimental validation. It provides significant insights into these variations in genes may affect biological functions as well as contribute to disease mechanisms, ensuring that these associations are validated [88].

3.2.6 SNAP2

SNP2 tool is an intended for examining whether single nucleotide polymorphisms (SNPs) alter the structure and function of proteins. SNP2 analyses whether a certain SNP is likely to be harmful or benign by examining the physicochemical features of proteins and the evolutionary conservation of amino acids [89]. This method integrates various sources of information through gene ontology annotations and genomic databases to enhance predictive precision. Its capacity to evaluate the probable pathogenicity of missense mutations is among its most critical responsibilities in identifying variants that may contribute to disease. SNP2 (Screening for Non-Acceptable Polymorphisms 2) program utilizes a support vector machine (SVM) algorithm to predict the functional consequences of amino acid substitutions on protein structure and function. Substitutions on protein structure and function. SNAP2 delivers values ranging from -100 to +100, where scores below zero are classified as harmful, while scores above zero are labelled neutral or benign [90].

3.3. Machine learning Approaches

Machine learning algorithms for SNP prediction incorporate the use of supervised and unsupervised models to find genetic variants related with diseases. Supervised algorithms such as Random Forest and Support Vector Machines are often used to categories SNPs based on their pathogenicity, whereas unsupervised approaches like clustering aid in detecting hidden patterns in genomic data.

3.3.1. Machine Learning-Based Data Collection

The SNP data was gathered from the Ensemble database, which serves as a public source. The raw dataset is made of MS Excel files. The text file comprises phenotyping information for the user and encodes SNP data in the VCF format [72].A VCF (variants Call Format) file has substantial variety and phenotypic information in the SNP data. Metadata such as the VCF format version and facts about the reference genome are supplied in the header lines at the start of a VCF file. variation is related with a single chromosome and indicates the precise nucleotide location at which the SNP is present on that chromosome. The reference allele represents the nucleotide discovered in the reference genome at that specific location. Each variant is allocated a distinct identification, usually obtained from databases such as

dbSNP. The VCF file includes alternate alleles that differ from the reference and a quality score indicating confidence in the variant call [91]. Each variation is granted a specific identity, sometimes acquired from databases such as dbSNP. The VCF file comprises alternative alleles that vary from the reference and a quality score expressing confidence in the variant call. Moreover, the VCF file contains crucial information, including the allele frequency and potential gene-related ramifications of the variation, as well as information regarding whether the variant has satisfied with requisite quality requirements. It offers genotyping data for datasets including several samples, identifying the presence of either reference or alternative alleles in each sample. This data offers complete genomic analysis [92].

3.4. Data Preprocessing

Data preprocessing is a significant step in the machine learning process. It is the phase of data preparation that raw data is and converts it into a more suitable format for the prediction. Preprocessing ensures that the data used provides more accurate predictions and increased reliance on the models [93]. In our study, we focused on selecting SNPs that appeared consistently across all genotyping datasets, ensuring a 100% call rate to maintain consistency. The technique used was based on different SNP variations form different suppliers, which might vary along parameters like the timing and the purpose of the tests. Consequently, Consequently, a total of more than 8,000 SNPs with overlapping effects identifed that could use for further study. SNPs relevant to the IFRD1-associated traits or diseases were retrieved from different bioinformatics databases. After verifying our dataset, it was determined only 45 SNPs met the criteria for inclusion, ensuring high-quality data for our analysis [94].The preprocessing mainly consists of the following important steps:

3.4.1. Data Cleaning

The process consists of identifying the SNP CVS file then the issues, such as outliers, duplicates, and missing values. The process of cleaning the SNP dataset guarantees that the model is trained with accurate and relevant data, which is the most important aspect of efficient prediction [95].

3.4.2 Data Transformation

SNP analysis requires data into proper format. This step consists of normalizing numeric values, converting the categorical variables into numbers for dictionary and scaling features which will bring uniformity to the entire dataset are the steps involved in this procedure [96].

3.4.3. Data Integration

Data integration in SNP datasets means combining different sources of data that relate to genetic variants. Through this, the target and features that can be used for model training are improved, which subsequently results in better analysis and higher precision of predictions[97].

3.4.4. Data Splitting

Our final phase involves the division of the preprocessed data into training and testing sets. In order to assess the effectiveness of our model, it is important that we maintain this distinction. The approach he uses guarantees that our model is robust and reliable when it generates predictions for SNP datasets[98].

3.5. Feature Selection

The method of feature selection was utilized to forecast the most essential feature on the SNP dataset. The process consisted of evaluating the correlation between all attributes and the scoring, applying a recursive feature elimination method with an XGBoost regressor and also doing logistic regression. These methods help to the identification of the most significant features that were the most revealing during the training and evaluation process of boosted decision tree models and logarithmic models respectively [99].

3.5.1. Random Forest

The Random Forest (RF) approach was utilized to produce a number of decision trees focused on the most important SNPs. Random Forest determines feature selection methods, evaluates the predictive importance of each feature, and thus that help in creation of more efficient decision trees [100]. The feature selection algorithms are essential for identifying the

characteristics that are most strongly correlation with the expressions of specific diseases, when high-dimensional data is used in the process of disease detection. The RF model was built up structured input matrix, in which (N) represents the total number of samples, and (X) and (Y) show the SNP values and the suitable output class for each sample, respectively. The model computed the out-of-bag error (OOBE), which estimates the prediction error for data points excluded from the training subset of all decision trees [101]. The SNP was subjected to a permutation procedure within the training data during the training phase, thus the model to determine the impact of that SNP on the accuracy of the predictions. The differences in OOBE were averaged and the resulting score was then normalized to give an accurate score for or the particular SNP[102].

3.6. Model construction

To predict an individual's susceptibility to SCA18, two machine learning models, RF-KNN and RF-SVM, were developed. Both models utilized Random Forest (RF) for feature selection to identify the relevant SNPs, which were then used as input for the KNN and SVM classifiers[103].

3.6.1. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a robust machine learning algorithm that is extensively employed in a variety of fields, including bioinformatics and genomics. It is particularly effective for the classification of high-dimensional data, such as SNP-based disease predictions[104]. Support Vector Machine (SVM) operates by identifying the optimal hyperplane that maximally separates data points from different classes, hence proving effective in scenarios when data is not linearly separable. Utilizing the "kernel trick," SVM can convert data into higher dimensions, allowing it to manage complex, non-linear patterns commonly found in genetic data [105].

3.6.2. k-Nearest Neighbors (k-NN)

The k-Nearest Neighbours (k-NN) approach is a basic instance-based learning mechanism that is applied in regression and classification. It categorizes a new data point by identifying its k closest neighbors by applying a certain distance metric, commonly Euclidean

distance, and determines the most popular class (for classification) or average value (for regression) of these neighbors as the prediction [106]. The selection of k significantly impacts performance: a small k may be susceptible to noise, while a larger k could reduce predictions but may obscure intricate patterns in the data. k-NN is extensively used in disciplines notably image recognition and genomics, identified for its simplicity and effectiveness in high dimensional data [107].

3.7. Protein Stability Analysis

To estimate the impact of mutations on IFRD1 protein stability, that we employed support vector machine (SVM)-based programs, such as I-Mutant2.0 [108] and Mupro [109], which utilize DDG (Delta Delta G) values and RI (Relative Importance) free energy to assess amino acid stability changes upon amino acid substitutions.

3.7.1 I-Mutant

I-Mutant 2.0 is a web server specifically designed for predicting the potential impact of non-synonymous single nucleotide polymorphisms on protein stability. The database will predict the changes the change in free energy (DDG) caused by alterations in amino acid residues within the protein sequence. A positive DDG value (> 0 kcal/mol) suggests an increase in stability, while a negative DDG value (< 0 kcal/mol) indicates a decrease in stability. Furthermore, through the prediction of stability caused by alterations due to this method allows researchers to select single nucleotide polymorphisms (SNPs) that are deleterious more focused studies of their investigations into their biological significance [110].

3.7.2 Mupro

Mupro is a user-friendly web-based tool designed to predict the effects of single nucleotide polymorphisms on protein stability. MuPro offers a simple interface where users can input the protein sequence and the specific mutation SNP to predict its potential impact on protein stability. Mupro is an online server developed to forecast the impact of single-site amino acid changes on protein stability. It utilizes Support Vector Machines (SVM) and Neural Networks (NN) to make predictions. Mupro score to indicate below 0 significant deleterious in protein stability [111].

3.8. Identification of Evolutionary Conservation

ConSurf tool was used to investigate the evolutionary conservation of amino acid substitution in IFRD1 protein. this tool needed a Fasta sequence as an input to determine the conservation status of the of the substituted residues [112].The result of the evolutionary conservation was created using an empirical Bayesian approach, and an amino acid's score ranged from 1 to 9, with 1 denoting changing amino acid and 7–9 representing evolutionarily conservative amino acids. This tool examines the degree of evolutionary preservation of amino acids at certain places within the protein's three-dimensional (3D) structure [113].

3.9. Functional consequences of point mutations

The functional repercussions of point mutations are vital in understanding how single nucleotide variations may affect protein function and lead to diverse disorders. These mutations may create missense, nonsense, or silent mutations, changing the protein's amino acid sequence, halting protein synthesis prematurely, or having impact on the protein’s structure, respectively [126]. For further inquiry the prediction of the structural effect of mutations on IFRD1 protein is done by employing project HOPE tool.

3.9.1. Project HOPE

Project HOPE is an advanced online tool designed to assess the structural and functional consequences of single amino acid substitutions in proteins [127]. It includes data from multiple sources, including three-dimensional structural predictions and sequence annotations from the UniProt database. By assessing user provided protein sequences, Project HOPE provides entire reports with vocal explanations, visual representations, and animations [128].

3.10. Protein Secondary Structure Analysis

Predicting protein secondary structure is crucial for understanding the structural and functional characteristics of proteins. The identification of local structural components, including α-helices, β sheets, and random coils, merely from a protein's amino acid sequence [114].

3.10.1. SOPMA

For secondary structure prediction analysis, we employed the SOPMA website which

combines a variety of algorithms, including Chou-Fasman.The tool's sophisticated algorithms and thorough grading system enabled us to achieve accurate predictions of alpha-helices, beta-sheets, and random coils. By estimating probability ratings based on sequence data, SOPMA created a formal framework for analysing the likelihood of various structure configurations [115].

3.10.2. NetSurfP

Netsurf was utilized for the prediction of a particular secondary structure and relative surface accessibility value of the amino acids[116]. To predict the secondary structure of the Wild Type and Mutant of the protein IFRD1, based on five different independent approaches. To it predict that amino acid relies on alpha-helix, beta-sheet and coils [117].

3.11. Protein Tertiary Structure Prediction

The unavailability of the three-dimensional (3D) structure of the IFRD1 presents an important challenge in examining its structural and functional properties. Predicted models provide valuable insights into understanding the structural consequences of mutations and the potential development of strategies for treatment utilizing different tools to predict the 3D structure of proteins [118].

3.11.1. Alpha fold

Due to the unavailability of the three-dimensional (3D) structure of the IFRD1 protein, computational approaches such as AlphaFold offer a strong solution for predicting its structure. AlphaFold, a system based on deep learning designed by DeepMind, has improved protein structure prediction by attaining near-experimental accuracy. Typically, the sequence is given in a FASTA format, which includes the amino acids in their appropriate order [119].AlphaFold utilizes this information Utilizing advanced algorithms and deep learning models to predict the three-dimensional (3D) structure of the protein. AlphaFold offers confidence in the score, including the pLDDT (predicted Local Distance Difference Test). A pLDDT score of 70 suggests accurate forecasting for specific places [120].

3.11.2. Swiss model

The Swiss model is an internet server that is employed to predict the 3 D structure of

proteins. Access to the Swiss model is given. is a regularly used web-based automated comparison model of 3-D protein structure and homology modeling server [121]. This service is aimed to facilitate protein modeling. The World Wide Web interface provides a variety of degrees of engagement in the Swiss model. To construct a three-dimensional model, the amino acid sequence of the protein may be supplied. Alignment, template selection, and model building are handled by the server [122].

3.12. Protein Structure Validation

Validation of predicted protein structures ensures their accuracy and trustworthiness for biological interpretations for study. Techniques such as Ramachandran plot analysis are utilized to analysis residue.

3.12.1. PROCHECK Server

TO predicted protein model was validated using the PROCHECK server, which conducts Ramachandran plot analysis to evaluate the model's stereochemical quality. It provides the number of residues in the favoured, allowed and outlier regions. If a significant number of residues are in the favored and allowed regions, the model is considered good. A significant percentage of residues in the favored and allowed regions indicates a robust and dependable model, confirming that its suitability for subsequent structural and functional analysis [123].

3.12.2. Discovery Studios

To evaluate the likely effect of nonsynonymous single nucleotide polymorphisms (nsSNPs) on the structure of the IFRD1 protein, we applied the Discovery Studios platform [124]. This effort employed Discovery Studios for showing molecular visuals and 3D protein structures, which permitted for interactive viewing and interpretation of molecular data. This platform facilitated the comparison of natural amino acids with their mutant equivalents and simplified the estimation of root mean square deviation (RMSD) values to highlight probable consequences and assess structural differences [125] .

3.13. Identification of Functional Domains

Identifying functional domains, which are specific regions in a protein that are responsible for particular biological processes, is an essential step in understanding protein function as well as the processes that are related to disease [129].

3.13.1 Conserved Domain Database (CDD)

The Conserved Domain Database (CDD) search tool used to identify functional domains and conserved regions among the protein sequence give information about the protein structure and function. By integrating domain models from databases such as Pfam and SMART, CDD facilitates evolutionary analysis and functional annotation. This approach is essential for understanding protein mechanisms and mutations prevalent in the disease[130][131].

3.13.2. InterPro

InterPro is a comprehensive database that integrates multiple protein annotation resources to classify protein families, domains, and functional sites. By incorporating data from databases such as Pfam, SMART, and PROSITE, InterPro facilitates the functional analysis of proteins across diverse species. It provides valuable insights into molecular mechanisms, supporting research in areas such as drug discovery, disease mechanisms, and evolutionary biology[132].InterPro will provide a thorough and comprehensive database for the functional annotation of proteins, facilitating researchers in comprehending the structural and functional characteristics of proteins across diverse species [133].

3.14. Molecular Dynamics Simulations

Desmond, software developed by Schrodinger LLC [134] was used to perform 100 nanoseconds (ns) Molecular Dynamics (MD) simulations, an accurately simulating the molecular motion according to Newton's laws of motion[135]. MD simulations were utilized to determine the interactions between the protein and the ligand, as well as the atomic-level conformational changes selected compounds, thus elucidating the process of binding to the target protein [136].The analysis of MD simulation, which included Newton's classical equation of motion, was utilized to predict the status of ligand binding in the physiological

environment. The A1 and A2 complexes were preprocessed using the default settings of the protein preparation wizard program. Maestro. Subsequently, the entire system was set up using the system builder tool [137]. By decreasing the energy of both systems and subsequently, allow them to equilibrate within the orthogonal box of the TIP3P (Intermolecular Interaction Potential 3 Points Transferable), a water model measuring (10Å×10Å×10 Å) with an OPLS_2005 force field was utilized as the solvent model [138].Throughout the simulation period, physiological conditions were simulated using 0.15M sodium chloride and 300K temperature and 1 atm pressure [137].Counter ions were utilized to neutralize the models. Trajectories were saved for inspection at intervals of 100 picoseconds (ps), and the Root Mean Square Deviation (RMSD) parameter was used to verify the stability of the protein-ligand complex over time [139].The MD simulation was successfully conducted using Dell T7810 with an Intel i5 -12th gen processor, 64 GB RAM, and 4070ti GPU on a system running.

Chapter: 04

Results And Analysis

4.Results and Analysis

This study highlights the main focus of the identification of harmful single nucleotide polymorphisms and their consequent alterations in the protein function and structural stability. The application of machine learning algorithms and used cross-validation technique for computational analyses its possible assessment of SNPs according to their predicted impact and disease progression. The molecular dynamic simulation was performed to evaluate the stability of the proteins provide the valuable insight into the modification of structure and molecular level.

4.1. Data Retrieval

The identification of the single nucleotide polymorphisms within the IFRD1 gene in depth examination used various bioinformatics tools and databases, such as Ensembl and dbSNP. A total of 20622 SNPs were identified analysis 5710 SNPs from non-coding regions and 528 from coding regions. The coding region SNPs was further filtered which resulted in the identification of 368 non-synonymous SNPs which lead to changes in the amino acid sequence and 160 synonymous SNPs that do not affect the amino acid sequence as shown in **(Figure. 4.1).** The retrieval and filtering of Single Nucleotide Polymorphisms data is a stepwise process illustrated in the diagram from first step Data Retrieval to use for Deleterious Variant Identification. Data for the IFRD1 gene is retrieved from a database the total number of SNPs in the IFRD1 gene is determined. Next Filters are used on this data, which leads to a decrease in the number of SNPs Used filtering based on SNPs analysis tools and specific potential deleterious identification of the IFRD1 gene.'

4.2. Machine Learning-Based SNP Identification

A total of 8,872 SNPs were identified for analyzing these datasets through the integration approach based on machine learning retrieved from the Ensembl database. Further refinement reduced this number to 45 potential candidates. The number of SNPs was then refined to 45 potential candidates identify the highest-ranked 22 SNPs were selected for in depth Analysis based on their significance function predicted pathogenicity, and potential impact on gene function.

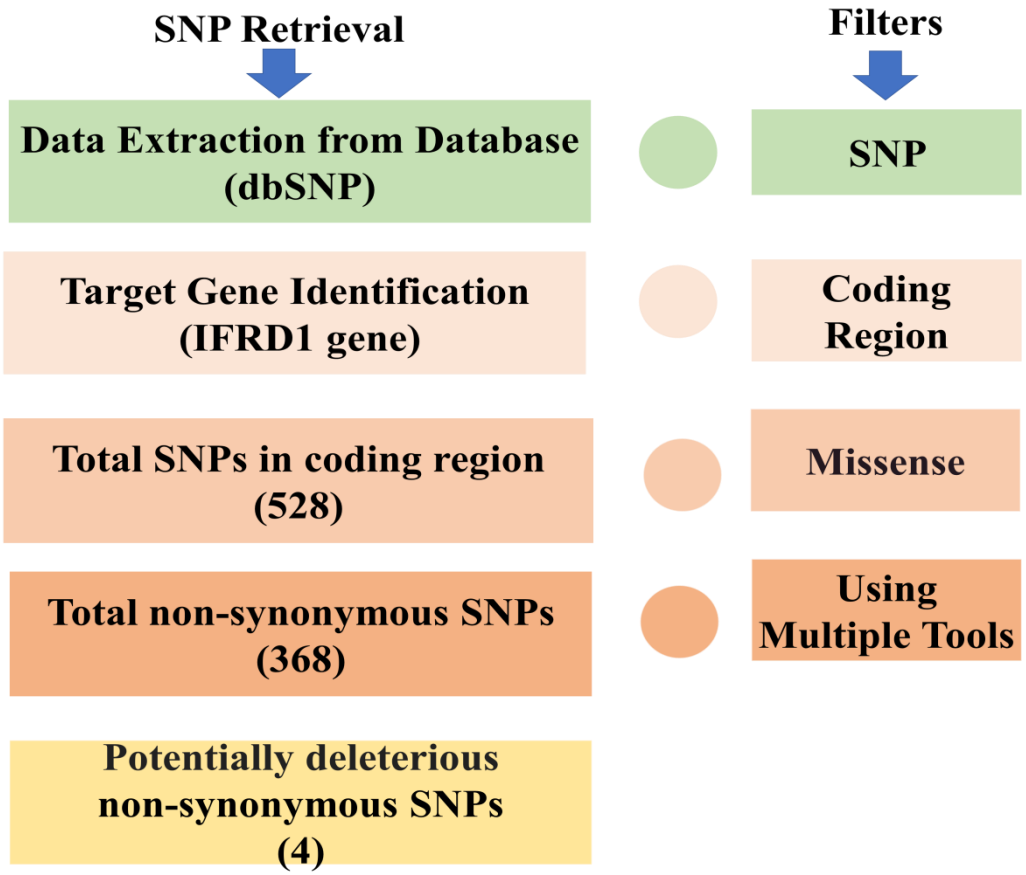


Figure 4. 1.SNP Analysis Workflow: From Data Retrieval to Deleterious Variant Identification.

The diagram depicts a process for retrieving and filtering Single Nucleotide Polymorphisms (SNPs) step by step. The process starts with retrieving data for IFRD1 gene from a database. Then, the total number of SNPs within IFRD1 gene is calculated. Filters are subsequently applied to this data, such as focusing on the coding region of the gene, which results in a reduced number of SNPs. Further filtering based on tools and potential deleterious effects refined the data set by identify specific SNPs of interest within IFRD1 gene.

4.3. Identification of Deleterious nsSNPs

To identify potentially deleterious SNPs that may substantially impact the structure or function of the IFRD1 protein, we employed a comprehensive *in-silico* analysis utilizing six distinct prediction tools: SIFT, PANTHER, PolyPhen-2, Meta-SNP, SNAP2, and SNPs&GO, polyphen-2. Our analysis revealed that four non-synonymous SNPs (nsSNPs) consistently exhibited deleterious predictions across all six computational methods, suggesting their potential to induce harmful effects on protein function.

The SIFT tool conducted an analysis of 368 nsSNPs from the IFRD1 gene where the prediction of their functional impact was based on the SIFT scores. The Variants scoring more than 0.05 were categorized as tolerant, while those with scores below 0.05 were considered as harmful. Additionally, SNPs not found in the database were categorized as not scored. As a result, SIFT identified 28 nsSNPs, including 10 damaging, 20 tolerant, and 320 unscored variants. To further assess the potential pathogenicity of these variants, the Panther tool evaluated 48 nsSNPs based on a threshold score of 0.5. Variants with scores equal to or greater than 0.5 were classified as possibly damaging, those exceeding 0.5 as probably damaging, and those below 0.5 as possibly benign. The analysis identified 16 nsSNPs as possibly damaging, 6 as probably damaging, and 4 as possibly benign. The functional impact of nsSNPs was further examined using the SNAP2 tool, which requires a protein sequence in FASTA format. Based on a default threshold score of 0, variants with scores greater than 0 were predicted to have a strong effect, while those with lower scores were considered neutral. The results indicated that 7 nsSNPs were predicted to have a functional impact, whereas 3 were neutral. SNPs&GO was utilized to determine whether nsSNPs were neutral or disease-associated by incorporating mutant location, substituting residue, and wild-type residue data. Variants with scores exceeding 0.5 were categorized as disease-associated, while those with scores below this threshold were classified as neutral. The tool predicted 4 nsSNPs as disease-related and 6 as neutral. PhD-SNP analyzed protein sequences along with mutant residue positions to classify variants based on a threshold score of 0.5. The results identified 5 nsSNPs as disease-related and 5 as neutral. Polphen2 tool analysis the score 0.9 identify 4SNP as probably damaging, 6 as possibly damaging. Meta-SNP, an integrated tool that combines multiple predictive algorithms, classified nsSNPs into four disease categories while identifying 6 variants as neutral. A detailed summary of these findings is presented in ***Table 4.1.***

The following four nsSNPs have been assessed for their potential impact on protein function

and disease association using multiple prediction tools: Among these tools, nsSNPs were considered deleterious, effect, disease, or probably damaging based on specific score thresholds: SIFT (score ≤ 0), PolyPhen-2 (score > 0.9), PANTHER (score > 0.5), Meta-SNP (score ≥ 0.5), SNAP2 (score > 0), and SNPs&GO (score > 0). The variant id- D179G: Aspartic acid (D) to Glycine (G) at position 179, R441Q: Arginine (R) to Glutamine (Q) at position 441, D308V: Aspartic acid (D) to Valine (V) at position 308, G287V: Glycine (G) to Valine (V) at position 287 as shown is **Table 4.2**.

4.4. Machine Learning Algorithms Utilized in SNP Mutation Prediction Tools

The machine learning algorithms employed in various SNP mutation prediction tools to assess the functional impact of genetic variants. The use of these algorithms improves the precision of predicting mutations linked to diseases and helps to identify possible biomarkers for a disease. This highlights numerous machines learning various machine learning approaches applied by different SNP mutation prediction tools each to particular algorithm assesses the and mutations affect diseases. The selection of these algorithms such as Naïve Bayes, Deep Learning, Random Forest, and Logistic Regression is based on the prediction targets. SNP prediction tools as illustrated in **table 4.3.** The following table a summary of the SNP prediction tool. This table is to provide extensive analysis of machine learning used for SNP mutation prediction tools. These tools to identify the functional consequences of mutation non-synonymous single nucleotide polymorphisms (nsSNPs), and identifying potential disease-associated variants.

Table 4. 1.Computational Prediction of nsSNPs and Their Functional Classification

S.NO	RSIDs	Amino Acid Change	SIFT	Polyphen-2	SNAP2	PANTHER	Meta-SNP	SNPs&GO	PhD-SNP
1	rs143002375	D179G	Deleterious	Probably. Damaging	effect	Probably Damaging	Disease	Disease	Deleterious
2	rs139029166	D59N	Deleterious	Probably. Damaging	effect	Possbaly. Damaging	Neutral	Neutral	Deleterious
3	rs182917954	R441Q	Deleterious	Probably Damaging	effect	Probably Damaging	Disease	Disease	Deleterious
4	rs141889729	K135R	Deleterious	Possbaly. Damaging	effect	Probably. Damaging	Neutral	Neutral	Neutral
5	rs200913560	D308V	Deleterious	Probably Damaging	effect	Probably Damaging	Disease	Disease	Neutral
6	rs151204901	K5T	Deleterious	Possbaly. Damaging	effect	Probably Damaging	Neutral	Neutral	Deleterious
7	rs370898862	G287V	Deleterious	Probably Damaging	effect	Probably Damaging	Disease	Disease	Neutral
8	rs145564103	R317T	Deleterious	Possbaly. Damaging	Neutral	Possbaly. Damaging	Neutral	Neutral	Deleterious
9	rs144614382	T74M	Deleterious	Possbaly. Damaging	Neutral	Possbaly. Damaging	Neutral	Neutral	Neutral
10	rs144427603	P2L	Deleterious	Possbaly. Damaging	Neutral	Possbaly. Damaging	Neutral	Neutral	Neutral

Table 4. 2.Potentially deleterious nsSNPs identified by six in silico tools.

S.NO	RSIDs	Amino Acid Change	SIFT	Polyphen-2	SNAP2	PANTHER	Meta-SNP	SNPs&GO
1	rs143002375	D179G	Deleterious	Probably. Damaging	Effect	Probably Damaging	Disease	Disease
2	rs182917954	R441Q	Deleterious	Probably Damaging	Effect	Probably Damaging	Disease	Disease
3	rs200913560	D308V	Deleterious	Probably Damaging	Effect	Probably Damaging	Disease	Disease
4	rs370898862	G287V	Deleterious	Probably Damaging	Effect	Probably Damaging	Disease	Disease

Table 4.3.Mutation predication tools

Tool	**Machine Learning Algorithm(s)**
SIFT	Position-Specific Scoring Matrix (PSSM), Empirical Rules
PolyPhen-2	Naïve Bayes, Logistic Regression
SNAP2	Neural Networks (Deep Learning)
PANTHER	Hidden Markov Models (HMMs), Logistic Regression
Meta-SNP	Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes
SNPs&GO	Support Vector Machine (SVM)
Mupro	Support Vector Machine (SVM)
I-Mutant2.0	Support Vector Machine (SVM), Neural Networks (NN)
ConSurf	Random Forest (RF)

4.5. Random Forest-Based SNP Analysis

The Random Forest algorithm was utilized to explore the handle large and high dimensional biologically data on analyzing significant variants, particularly those linked to IFRD1 gene and its potential role in association with Spinocerebellar Ataxia Type 18 (SCA18). The application of machine learning techniques to investigate well organized and evidence-based method for SNP prioritization, which in help to identify pathogenic variant linked to the neurodegenerative diseases SCA18.Furthermore the Random Forest was used to identify key attributes significant characteristics as well as reduce data dimensionalities of the dataset Prior to the SNP analysis. The Random Forest can be use with feature selections identify important SNPs prediction associated with disease progression. The RF model showed high performance, score accuracy 62.9%, precision 67.8% and sensitivity 81% respectively. These findings, as illustrated in **Figure 4.3**.

4.5.1 Random Forest for SNP Selection

Radom Forest classifiers are an efficient approach for identifying significant SNPs, providing it a valuable tool for feature selection beyond classification performance. This approach facilitates comprehension of SNP data by identifying significant genetic changes with substantial predictive importance. The robustness of RF is identified as critical features, remaining unaffected by to alterations in the dataset. Significantly, 22 SNPs, distinguished by allele changes from A to G and G to A/C, highlight the complexity of genetic variation. SNPs exhibiting distinct allele shifts might reveal novel insights into disease associations when determined by robust RF machine learning algorithms underscore the capability of RF in optimizing SNP analysis and improving the comprehension of IFRD1-associated pathogenic variants in SCA18 disease. The SNPs, demonstrating particular allele changes from underscore the complexity of genetic variation and emphasize the effectiveness of Random Forest in determining significant variants associated with disease susceptibility. As shown in **Figure 4.4.**

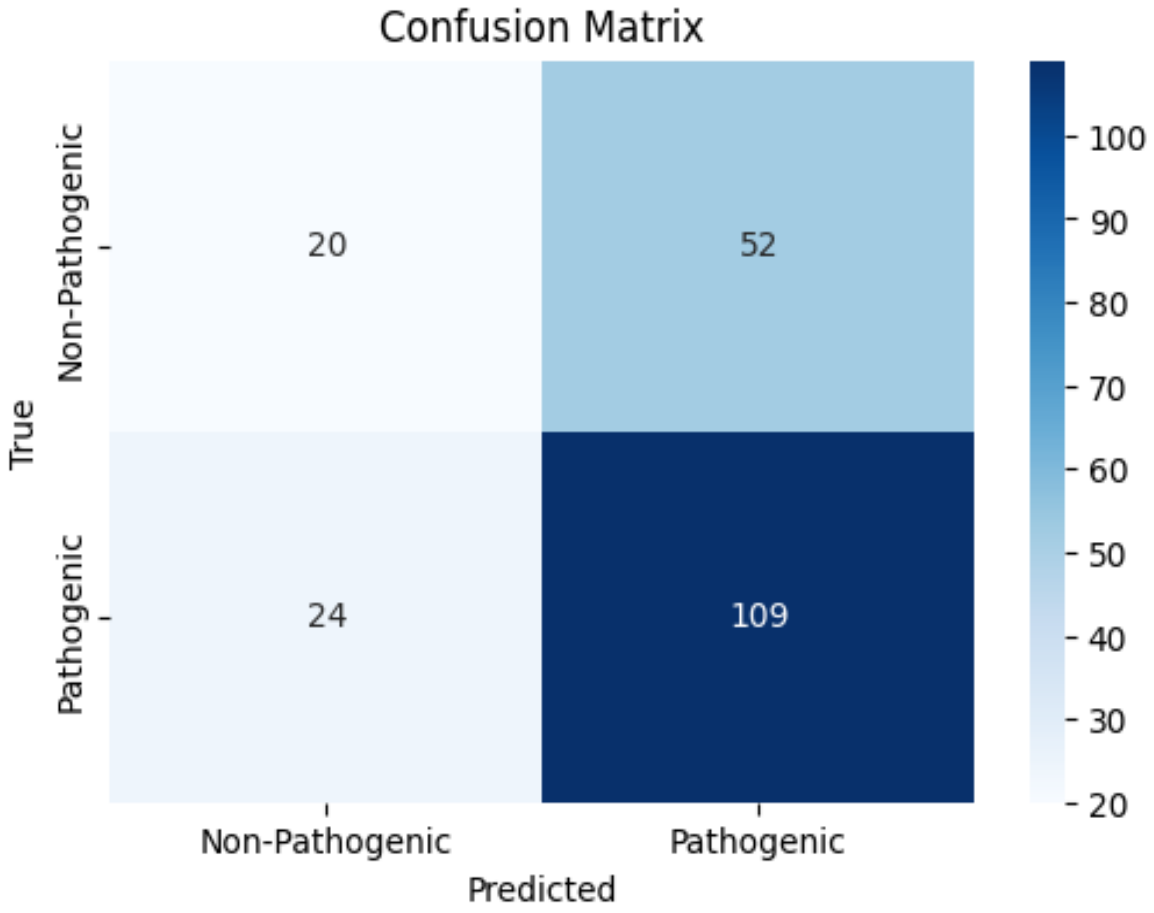


Figure 4.2 .Identification of SNPs

This figure shows the confusion matrix of the Random Forest (RF) model used for classification. The RF model achieved an accuracy of 62.9%, precision of 67.8%, and sensitivity of 81%, indicating strong classification performance.

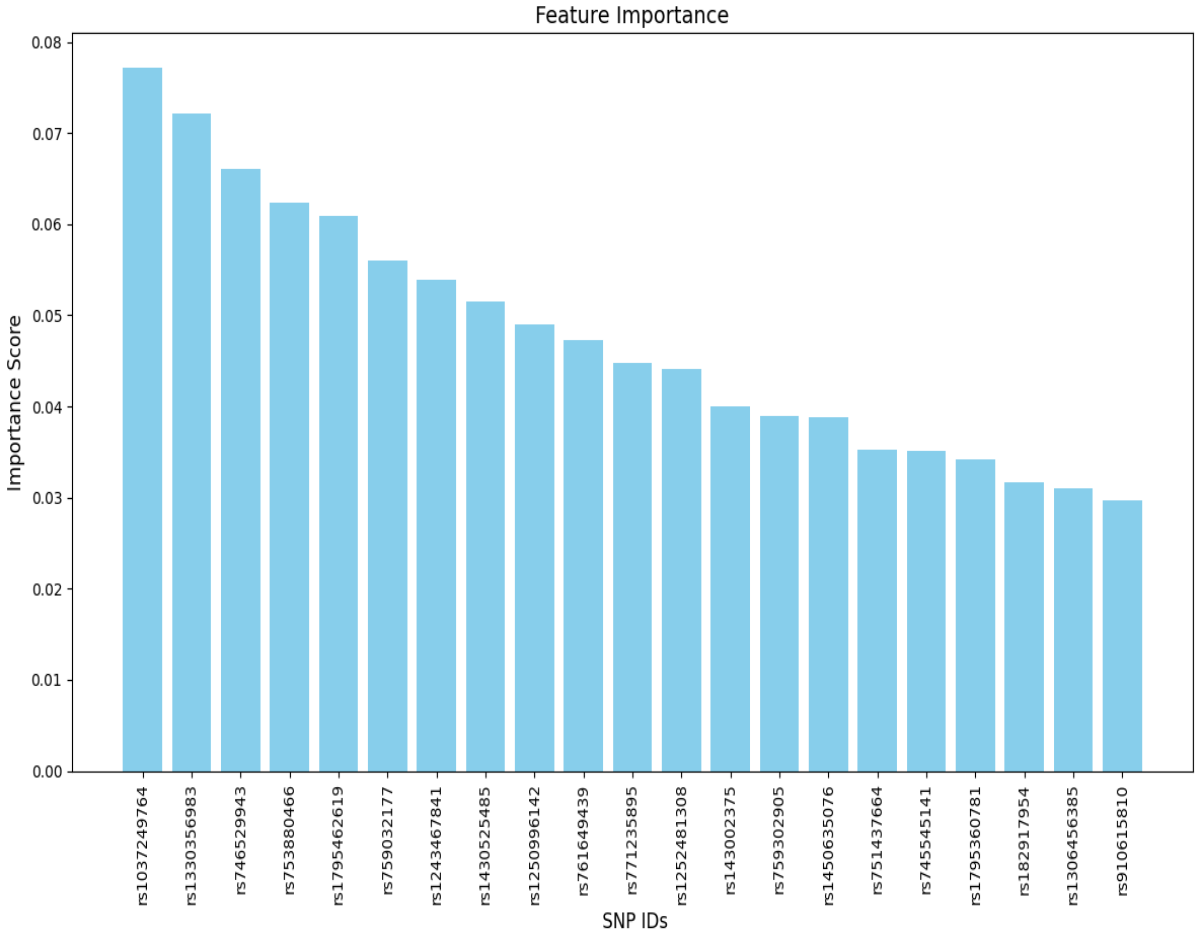


Figure 4. 3.Random Forest-based selection of SNPs for disease prediction.

This figure shows the identification of important SNPs based on feature importance scores generated by the Random Forest model, highlighting variants with a stronger contribution to disease prediction.

4.6. Model Comparison

The SVM and kNN algorithms were employed on the SNP datasets to build the baseline models. after that the classifiers analyzed for the SVM and kNN of 22 IFRD1 SNP associated with SCA18. The Random Forest was applied as feature selection method when combined with Support Vector Machines and k-Nearest Neighbors. The RF-SVM and RF-kNN was assessed in comparison with the baseline models. The findings indicated that the integrated models obtained enhanced classification performance comparable to the baseline models.

4.6.1. RF-SVM (Support Vector Machine

The Support Vector Machine (SVM) is a robust supervised machine learning algorithm used frequently for classification purposes. It operates by identifying the optimal hyperplane that best separates data points into distinct classes. Support Vector Machine (SVM) was initially employed as the baseline model for identifying SNPs of the ifrd1 gene linked with Spinocerebellar Ataxia Type 18 (SCA18). To improve the overall performance of the model, Random Forest was used for feature selection, and it was integrated into the RF-SVM model. The RF-SVM model gives better results show an obtained an accuracy of 64.8%, a precision of 64.8%, and a certification rate of 81% respectively. The combined models demonstrated in this **figure 4.5** provide an important in the overall model efficacy in the SNP analysis through the use of RF-SVM classification techniques.

4.6.2. RF-kNN (k-Nearest Neighbors)

The RF-kNN model is a combination of RF and k-NN classifiers. By using the RF-based for feature selection capability and KNN's effectiveness in identification the pathogenicity .In this supervised machine learning method, first Random Forest is for identifying and evaluate the most significant features from the SNP data and kNN classifier to predict the pathogenicity of genetic alteration the RF-KNN show a good score RF-KNN modelhas precision, sensitivity, and accuracy of 66.6%, 78%, and 60.4% respectively in its overall classification as shown **Figure 4.6**.

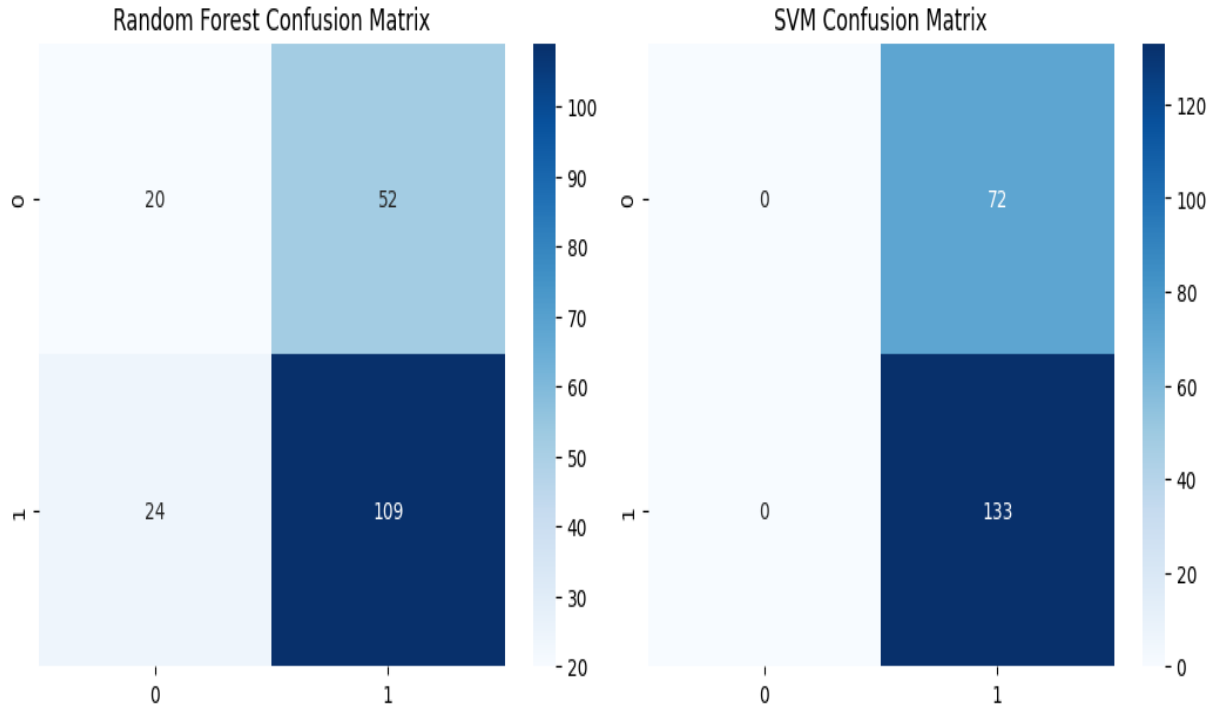


Figure 4.4. RF-SVM model improves SNP classification accuracy for SCA18

This figure shows the improvement in SNP classification for SCA18 using the combined RF-SVM model. The model achieved an accuracy of 64.8%, a precision of 64.8%, and a sensitivity of 81%, demonstrating enhanced performance in disease prediction.

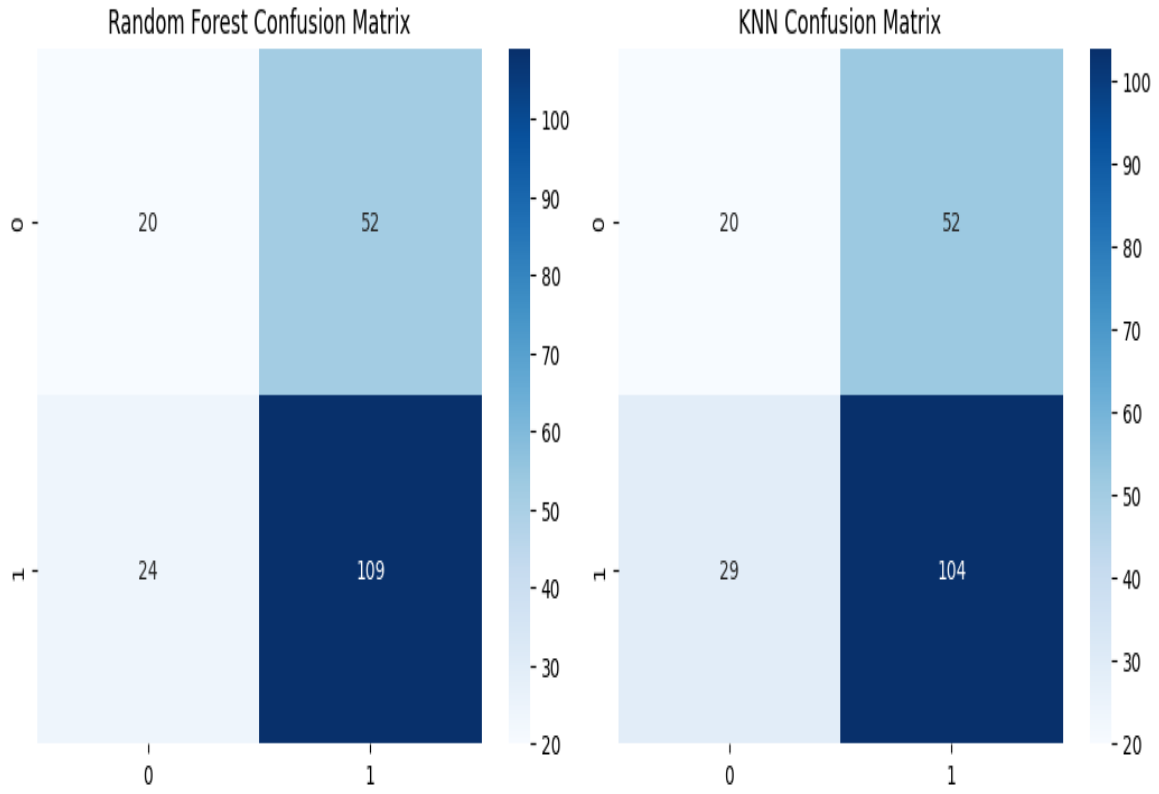


Figure 4.5. Efficacy of the RF-kNN model in detecting harmful variants

The Random Forest–k-Nearest Neighbors (RF-kNN) hybrid model achieved an overall classification accuracy of 60.4%, a precision of 66.6%, and a sensitivity of 78%. The figure highlights the model effectively identifies disease-associated SNPs, demonstrating its ability to distinguish harmful variants and providing insight into its potential genomic variant analysis for SCA18.

4.7. SNP-Based ML Model Evaluation via AUC

In this study, Machine learning techniques were employed in predicting individual susceptibility to Spinocerebellar Ataxia Type 18 (SCA18) based on single-nucleotide polymorphism (SNP) data. The identification and selection of significant SNPs linked to SCA18, followed by the application of comprehensive classification methods to improve accurate predictions. The RF-SVM model exhibited superior performance among the examined models, exhibiting higher accuracy, precision, sensitivity, and area under the curve (AUC) compared with the RF-kNN and baseline models. The Random Forest (RF) model achieved an AUC score of 0.72, while the RF-SVM model obtained score of 0.44, and the kNN classifier model AUC of score 0.64. These findings emphasize the importance of using different machine learning algorithms in analyzing complex SNP datasets consequently improving these associations between the genotype and predictive modelling in SCA18.This further demonstrates that the combination of machine learning approaches employed in proposed models may effectively forecast susceptibility to SCA18 with commendable efficacy, despite the few genetic variables included in the model development, as shown **in Figures 4.7,4.8,** and **4.9.**

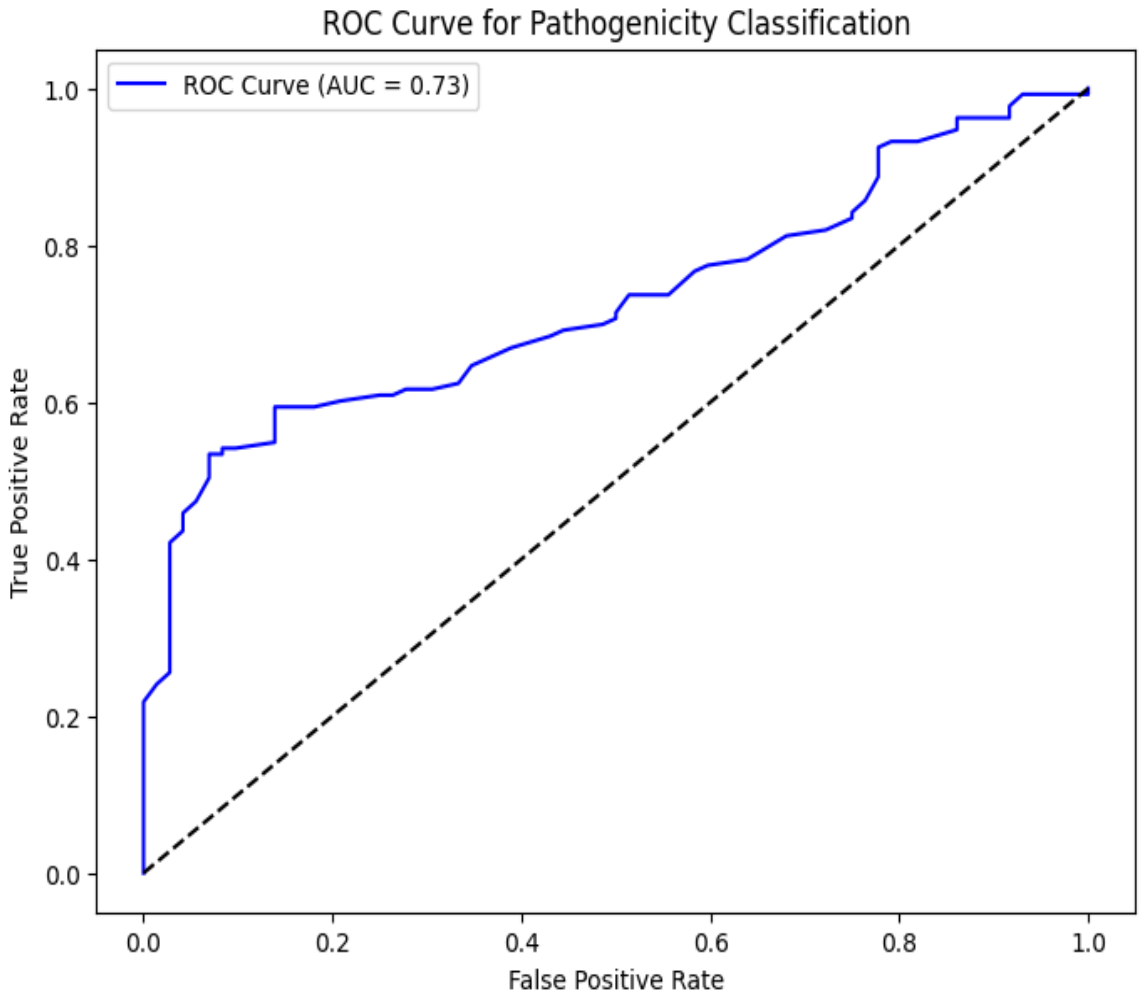


Figure 4.6.RF model achieved an AUC of 0.72, demonstrating its efficacy in SCA18 SNP classification

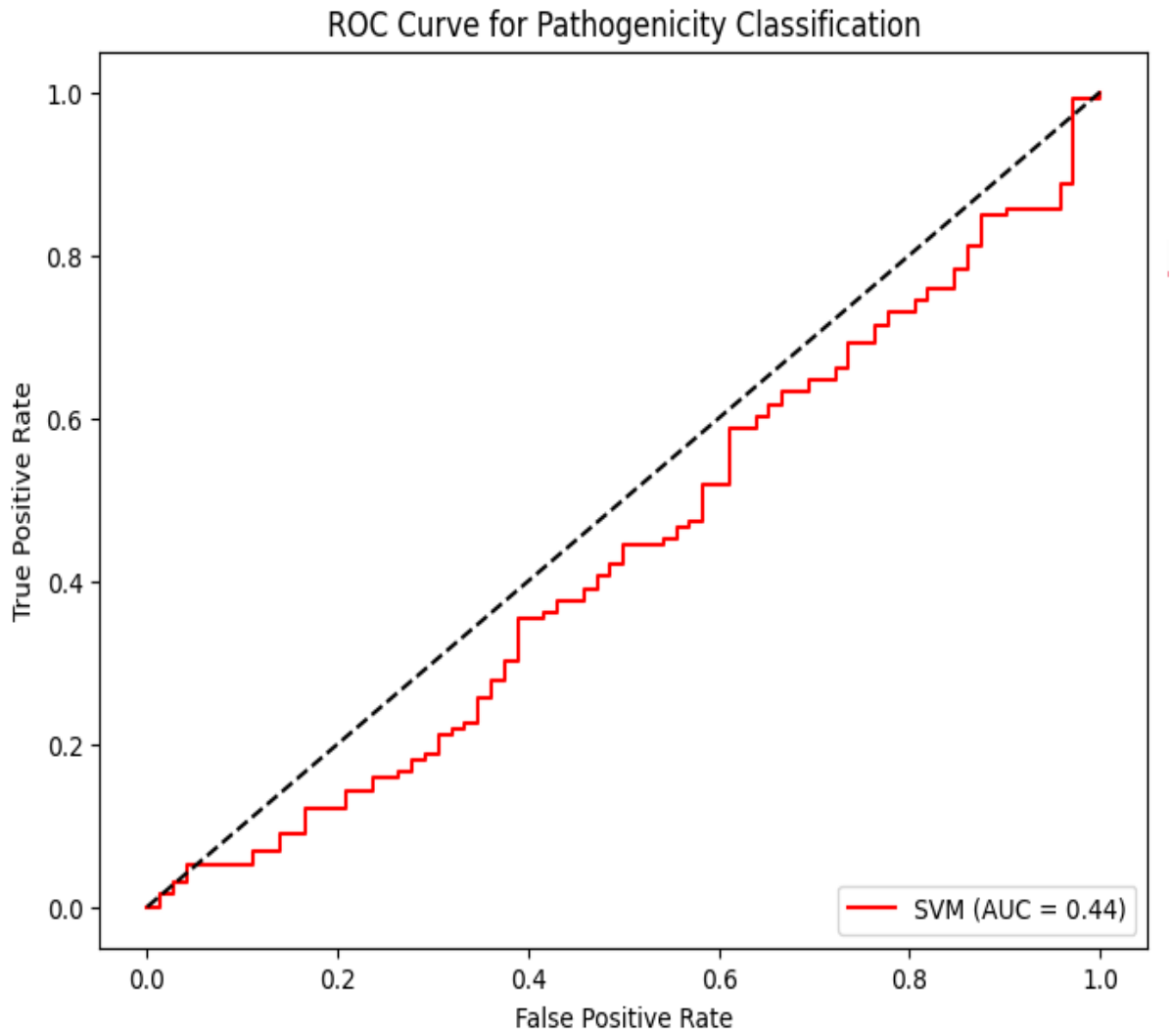


Figure 4.7.RF-SVM model achieved 0.44 AUC, highlighting its impact in SNP prediction

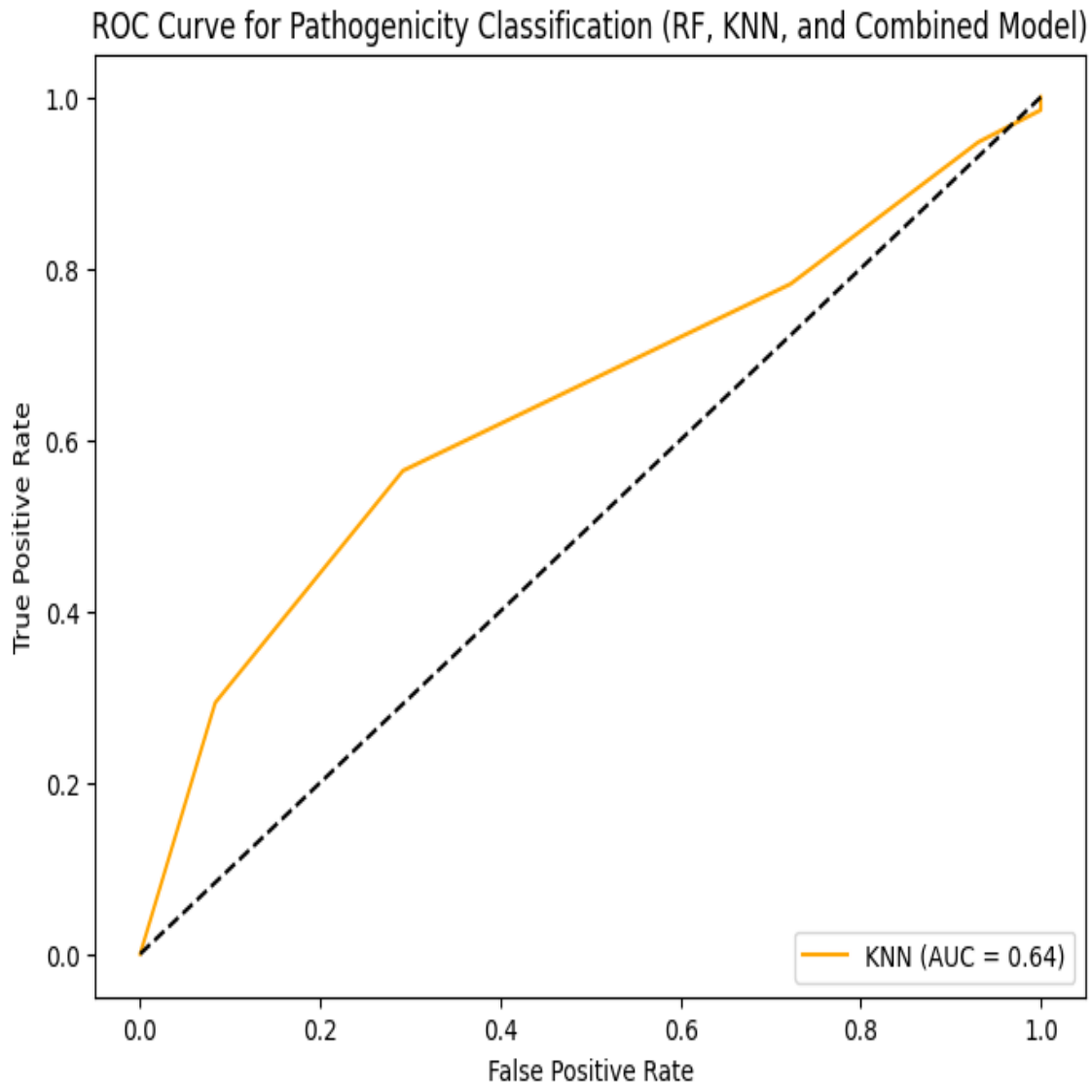


Figure 4. 8.kNN model achieved 0.64 AUC, highlighting its role in SCA18 classification

4.8. Identification of Protein Structural Stability

The functional analysis tools were first used to identify significant SNPs. Subsequently, cross-validation techniques were employed to enhance the selection, reducing it to the top 22 SNPs. Two specific variant IDs rs771235895 and rs1252481308 are included to provide a thorough examination, despite their potential impact on stability. A common variation ID rs182917954, was identified in both the functional analysis and cross-validation steps, elevating the overall consistency of the outcomes.

A detailed computational analysis utilizing I-Mutant and Mupro tools was conducted to systematically investigate the effects of point mutations on the structural stability of the IFRD1 protein. Our study revealed 368 nsSNPs as possibly deleterious. Among the four potentially harmful non-synonymous SNPs (nsSNPs) identified in the IFRD1 gene **(Table 4.3),**

our comprehensive analysis revealed that two specific variants, rs143002375, rs771235895, rs1252481308 and rs182917954, exhibit a significant destabilizing effect on the protein structure, leading to a substantial decrease in protein stability, as demonstrated by the data presented in **Table 4.3**. The selection criteria of these variants were based on the statistical parameters, as described in the methodology section.

Table 4.3, provides the outcomes of forecasting investigations using with I-Mutant and Mupro. Four non-synonymous single nucleotide polymorphisms (nsSNPs) were identified as potentially deleterious, affecting protein stability or functionality. In the I-Mutant predictions, the Reliability Index (RI) represents the confidence level, with values around 9 denoting higher levels of reliability. In Mupro, the ΔΔG-free energy change (kcal/mol) measures the change in stability, with lower values indicating decrease stability.

Table 4.4.Effects of nsSNP on protein stability determined by I-mutant and Mupro

S. no	RSIDs	Amino Acid Change	I-Mutant	RI	Mupro	DDG-Free energy change value (kcal/mol)
1	rs143002375	D179G	DECREASE M/G	9	DECREASE	-0.91678812
2	rs182917954	R441Q	DECREASE T/Q	8	DECREASE	-1.111973
3	rs771235895	R396Q	Decrease (A/Q)	8	DECREASE	-1.1422887
4	rs1252481308	D331G	Decrease(V/G)	9	DECREASE	-0.54045494

4.9. Identification of Functionally Conserved Residues

We analyzed and investigate the evolutionary conservation region of amino acid the wild type IFRD1 protein using the ConSurf. The IFRD1 protein consist of 451 amino acid the evolutionary conservation identified nsSNPs located in high-risk positions, which could have a very strong impact on the protein structure function IFRD1 protein. ConSurf analysis predicted that the four nsSNPs R441, D179, D331G, and R396Q protein are located in evolutionarily conserved regions was identified as functionally significant, while structural importance. This finding suggests this amino acid important role for the protein function and valuable impact for genetic variations**. Figure 4.10** shown that these SNPs were located in the highly converted region.

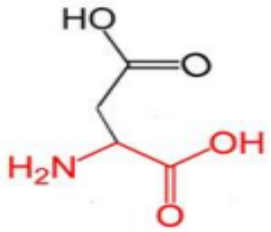
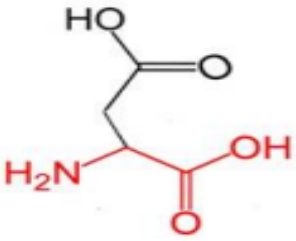
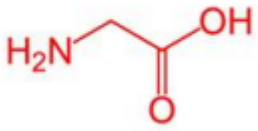
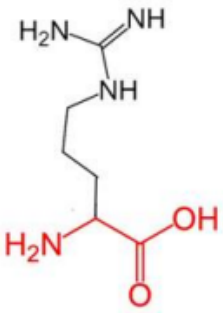
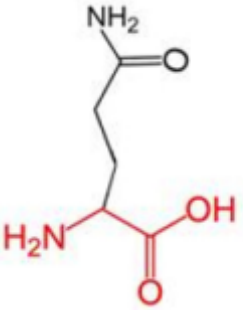
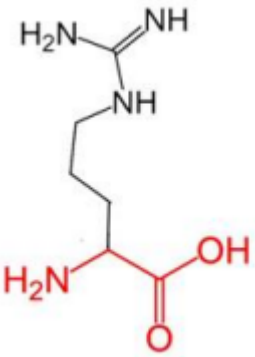
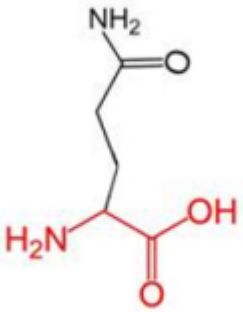
4.10. Structural Consequences of Point Mutations on the Protein

Single nucleotide polymorphisms effect on the IFRD1 protein were evaluated using the project HOPE analysis. This tool assesses each mutation amino acid substitutions impact of on specific position on the structure of the IFRD1 protein. The pathogenic non-synonymous single nucleotide polymorphisms (nsSNPs) that we explore the consequence of the impact on the amino acid size, charge, and hydrophobicity. Four mutations were identified leading to a reduced amino acid size. In the four nsSNP decrease the amino acid size and charge these specific mutations region D331G and D179G which changed the charge from negative to neutral, and R441Q and R396Q mutations region were found change the charge form positive to neutral. The mutation in the amino acid possible changes of the protein structure, and interactions with other molecule effect the overall protein function., as shown in the **table 4.7**. In this study focused on the non-synonymous single nucleotide polymorphisms present in the IFRD1 protein and their pathogenicity. The table highlights the impact of mutation on the properties of amino acids, size, charge, and hydrophobicity, as well as the structural modification in the protein.

Figure 4.9.ConSurf analysis of IFRD1 reveals conserved amino acids with evolution.

Evolutionary Conservation of the IFRD1 gene demonstrated by Consurf. This diagram displays the outcomes obtained from using the ConSurf tool, where conservation scores are represented by nine distinct color codes. These color codes signify the evolutionary relationships observed among sequence homologs. Notably, the variants R441Q, R396Q and D179G, D331G exhibited a conservation score of nine, indicating a high level of conservation.

Table 4.5. Results of the Project Hope Analysis of Structural and Functional Parameters for Wild-Type and Mutant Proteins

		Wild-Type IFRD1	Mutant IFRD1
D179G	**Amino acid structures**		
D331G	**Amino acid structures**		
R396Q	**Amino acid structures**		
R441Q	**Amino acid structures**		

4.11. Secondary Structure Prediction Analysis

Analyzing secondary structures is particularly important when investigating the effects of SNPs as the arrangement of α-helices, β-sheets, and random coils is crucial for protein stability, folding, and activity [140]. Each of these structural elements contributes uniquely to protein function such as α-helices are essential for enzyme activity and membrane function, β-sheets provide structural support and facilitate protein-protein interactions, and random coils are important for protein folding, signaling, and molecular recognition. Mutations within these regions can have profound effects, leading to protein misfolding or aggregation, which can result in neurodegenerative disorders [141]. Consequently, analysis of IFRD1's secondary structure is crucial for identifying SNP-induced changes and elucidating the complex relationships between protein structure, function, and disease mechanisms.

To understand the structural consequence of the SNPs on IFRD1, the wild-type protein and mutant type IFRD1 Proteins and their variants secondary structures were analyzed using the SOPMA server. As a result, the predicted secondary structure of IFRD1 revealed a distribution of α-helices, β-sheets, β-turns, and random coils. The wild-type IFRD1 protein, consisting of 451 amino acids, was composed of 259 residues (57.43%) in α-helices, 16 residues (3.55%) in β-sheets, and 177 residues (39.25%) in random coils as shown in **Table 4.6.** Variants exhibited distinct secondary structure profiles: for the mutant protein, the structure included 259 residues (57.65%) in α-helices, 14 residues (3.10%) in β-sheets, and 168 residues (37.25%) in random coils. Additionally, two SNPs, D331G and D179G, impact the conformation, whereas the other SNPs influence the extended strand structure. The two SNPs, R396Q and R441Q, specifically induce an alteration to random coil, while four SNPs affect the structure from extended strand to random coil. Graphical analysis of secondary structure elements **(Figure. 4.10**) revealed that mutations at positions 179 ,331(D to G in a β-sheet region) and 441,396 (R to Q in a coil region) caused alterations in secondary structure, potentially disrupting IFRD1's structural integrity and function. The wild-type protein shows a more stable and consistent distribution of α-helix and β-sheet elements across the sequence. The mutant type exhibits noticeable fluctuations with increased coil and turn regions, indicating that amino-acid substitutions disrupt secondary structure stability. Our analysis reveals that SNP-induced alterations play a critical role in disrupting IFRD1's stability and function. our prediction that the identified SNPs have a profound impact on the protein.

Table 4.6. Comparison of Secondary Structure Composition Wild Type and Mutant Type IFRD1 protein

Protein	**Secondary Structure**	**Total Residues**	**Residue (%)**
Wild type (IFRD1)	Alpha helix(H)	259	57.43%
	Beta sheet(E)	16	3.55%
	Random Coils	177	39.25%
Mutant IFRD1	Alpha helix(H)	269	59.65%
	Beta sheet(E)	14	3.10%
	Random Coils	168	37.25%

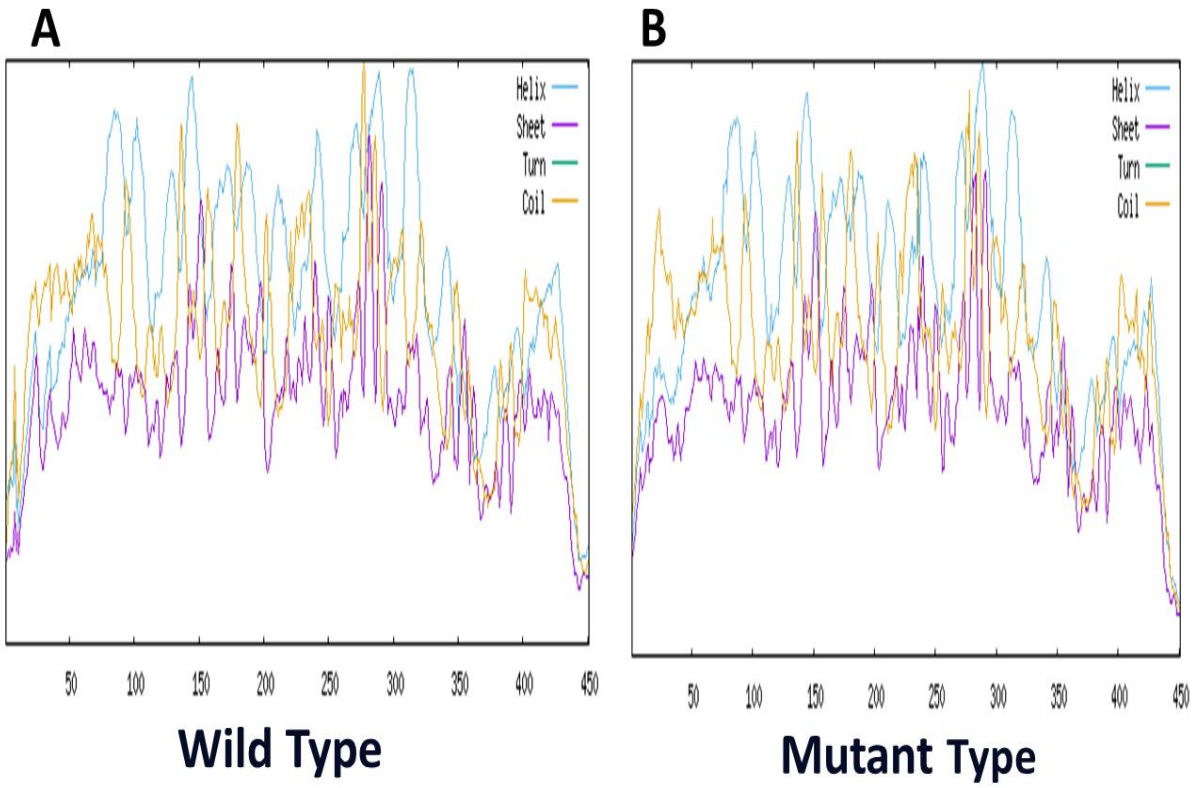


Figure 4.10.Predicted Secondary Structure of the Wild-Type IFRD1 Protein and Mutant type.

The graph, generated using the SOPMA server, illustrates the predicted secondary structure elements along the amino acid sequence (X-axis), with probability scores for each structure type (Y-axis). Alpha-helices are represented by light blue lines, beta-sheets by purple lines, beta-turns by green lines, and random coils by orange lines. Notable mutations at positions 179 ,331(D→G in a beta-sheet region) and 441,396 (R→Q in a random coil region) result in alterations to the secondary structure, highlighting potential functional impacts of these variants**.**

4.12. Homology Modelling

Due to the unavailability of the three-dimensional (3D) structure of the IFRD1 protein we retrieve the 3D structure of the IFRD1 protein Alpha fold. The protein sequence was used to generate three additional protein models using the Swiss Model. The Swiss Model Utilizes multi-template homology modeling to accurately predict the three-dimensional structure of proteins. Template selection was based on the highest percentage of sequence identity and query coverage. Swiss-Model ID O00458.1A, with GMQE values of 0.84, shows sequence identities of 100%, 99.78% with the IFRD1 protein, which comprises 451 amino acids. The multi-template homology modeling successfully generated a 3D model of the target protein through sequence alignment with the template structures, by using Discovery Studio, we performed a comprehensive analysis of IFRD1 proteins to determine their structural and functional characteristics. This analysis provides significant insights into the effects of nsSNPs on IFRD1 protein as detailed in **table 4.7**.

Ramachandran plot is used for validation to evaluate the quality of a protein model by verifying that its dihedral angles are within allowed conformational regions. This helps confirm that the protein structure is both realistic and accurate. The best-predicted models for wild-type and mutated IFRD1 proteins were evaluated using the Ramachandran plot to ensure the highest scores, 86.0%, 86.4%, detailed in **(Table 4.8)**. This analysis confirmed that the models exhibited favorable stereochemical quality and structural accuracy, as depicted in **(Figure 4.11)**

The Ramachandran plots show that the wild-type IFRD1 protein has a stable structure, with 86.0% of its residues in favored regions and only 2.0% in disallowed regions. The variants mutant IFRD1 maintain a similar overall distribution, with about 86.4of residues in favored regions and 2.0% in disallowed areas. However, the mutant displays localized structural deviations, implying that small modifications may influence the protein's function while maintaining its overall structure. The Swiss model which utilizes multi-template homology modeling to predict the 3D structure accurately. The Ramachandran plot was used to validate the protein structure.

Aspartic acid (Asp/D) and arginine (Arg/R) are two important amino acid that have that play substantial roles in function various biological and cellular activities. Asp is involved in the processes of protein synthesis, energy metabolism, and neurotransmission, while Arg is a

major role in protein synthesis, cell signaling, and immune responses [142] . Notably, Arg has a unique chemical structure and a well-established safety profile, making it an effective solubilizing and refolding agent for proteins. The IFRD1 protein, which contains both Asp and Arg residues, is affected by specific mutations. The predicted single nucleotide polymorphisms (SNPs) in the wild-type and mutated IFRD1 proteins involve these critical residues at specific position, linking them to pathologically significant diseases **Figure 4.12.** However, further investigation is necessary to determine the specific influence of these mutations on disease mechanisms. The highlighting mutant type protein positions of the mutations in the IFRD1 protein at the position R441Q, R396Q (Arginine to Glutamine), illustrating the structural changes and potential pathological effects and the position D179G, D331G (Aspartic acid to Glycine) demonstrating the alterations in the protein structure and associated disease implication.

4.13. Prediction of nsSNPs in IFRD1 Protein Domains

The InterPro, Pfam, and NCBI Conserved Domain Search tools were used to identify two important domains of the IFRD1 protein. The C-terminal domain consists amino acids 392–445, whereas the N-terminal domain consists of 50–347amino acids. The central region of the IFRD1 domain, spanning amino acids 347–392, is situated between both of these regions. In the N-terminal domain, single nucleotide polymorphisms (SNPs) D179G and D331G are identified, whereas R396G and R441G are situated in the C-terminal domain. The structural integrity and functional interactions of these domain-specific SNPs may influence IFRD1's role in cellular processes and disease pathways. This may lead to alterations in cellular signaling and transcriptional control. These mutations may be associated with an increased risk of disease development, highlighting the significance for more investigation into the functional consequences of IFRD1 mutations. The domain shown as **Figure 4.13.** *The* highlighting alterations in the N-terminal and C-terminal domains. The SNPs D179G, D331G, R396G, and R441G may alter cellular processes and contribute to disease development by altering protein function.

Table 4.7. The protein Structure predication was conducted using the Swiss Model

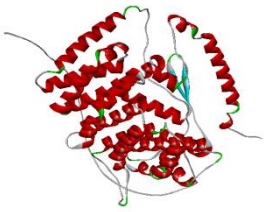
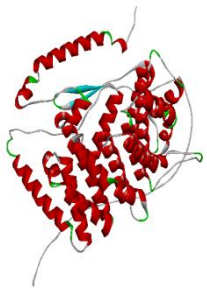
S.NO	Structure	Prediction	Sequence identity	GMQE-value
1		Wild type IFRD1 Alpha fold	100%	
2		Mutant type IFRD1 Swiss model	99.56%	0.84

Table 4. 8. Allowed and disallowed regions in Ramchandern plot

Protein Model	Favoured regions [A, B, L] (%)	Additional Allowed Region [a, b, l, p] (%)	Generously Allowed Region [α, β, l, p] (%)	Disallowed Region (%)
Wild-Type IFRD1	86.0%	9.3%	2.7%	2.0%
Mutant IFRD1	86.4%	9.1%	2.5%	2.0%

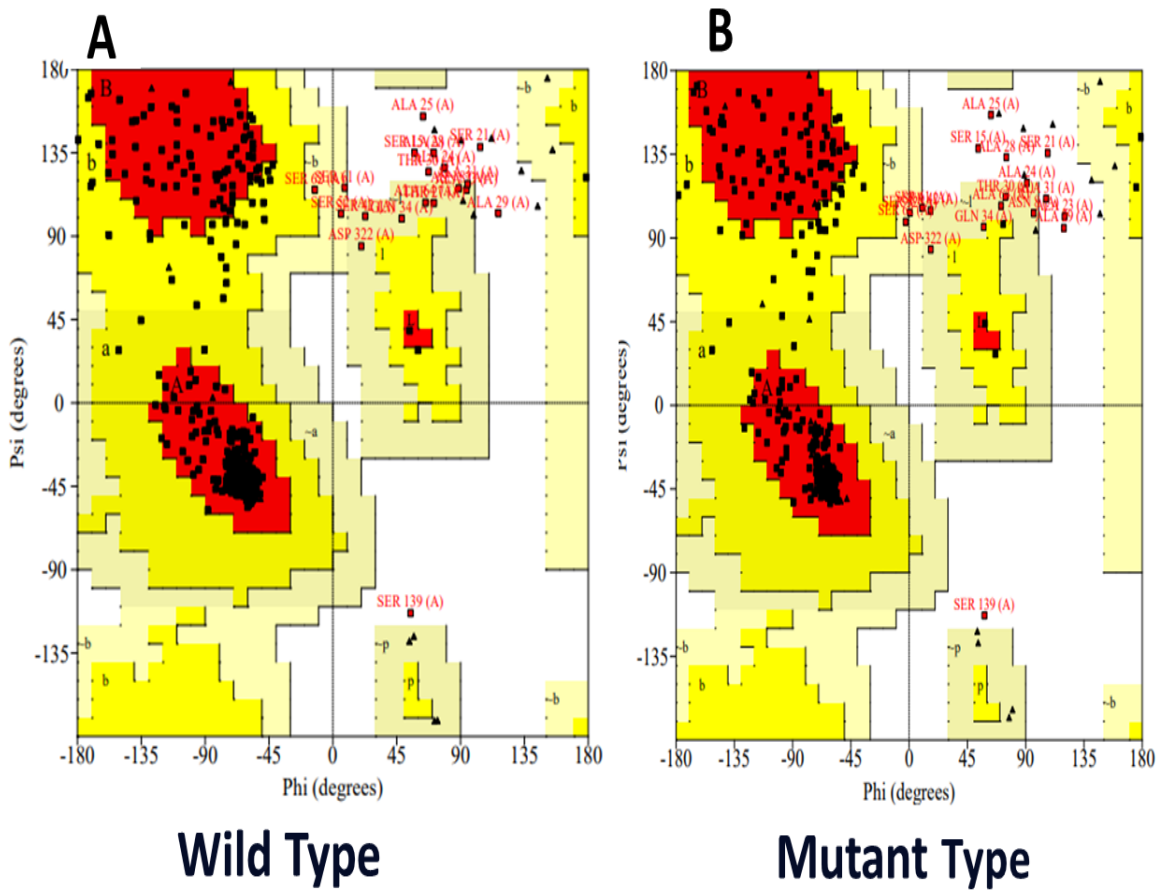


Figure 4.11. The characterization of mutations in the IFRD1 protein using the Ramachandran plot

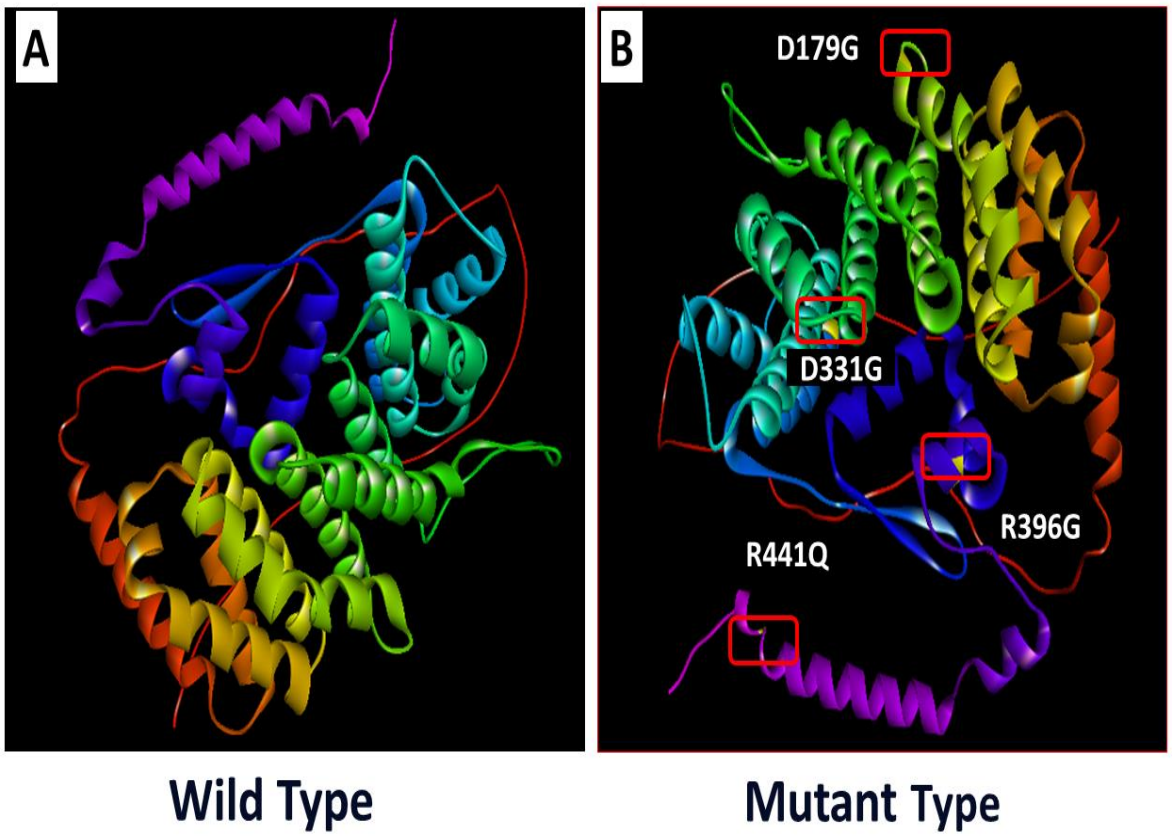


Figure 4.12. Structural Impact of IFRD1 Mutations. A) wild-type IFRD1 protein and B) Mutant type protein.

The figure Show the (A) wild-type IFRD1 protein, highlighting positions of the mutations in the IFRD1 protein at the position (B) mutant type protein position 441,396 (Arginine to Glutamine), illustrating the structural changes and potential pathological effects and the position 179 ,331 (Aspartic acid to Glycine) demonstrating the alterations in the protein structure and associated disease implications.

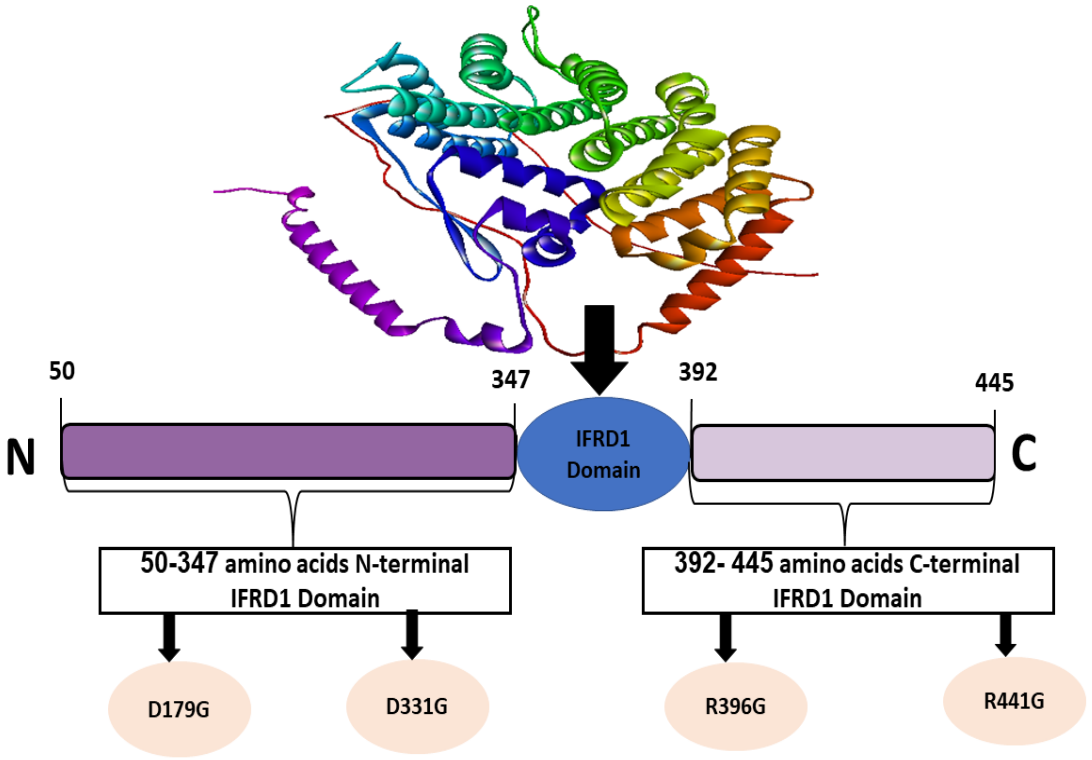


Figure 4. 13.Predicted SNPs within the IFRD1 protein domains

The Figure Show Structure of IFRD1, highlighting the N-terminal region of the protein, zinc finger domains, and the C-type carboxy-terminal tail

4.14. Molecular Dynamic Simulations

To further elucidate protein, we performed MD simulations after assessing stability and evaluating energetic contributions. For this purpose, each protein wild type IFRD1, and mutant IFRD1 protein was subjected to 100 nanoseconds of MD simulations to compare conformational changes induced by mutations and comprehensively understand mutation-induced effects. The simulation trajectory was comprehensively analyzed through the examination of several key parameters, including Root Mean Square Deviation (RMSD), Root Mean Square Fluctuations (RMSF), secondary structure elements (SSE), and the radius of gyration (RG) of the protein. These parameters were systematically evaluated throughout the entire time duration of the Molecular Dynamics Simulations (MDS), providing a detailed understanding of the protein's behavior and dynamics.

4.15. Wild type and Mutant Protein Structure Stability Analysis

The analysis of the wild-type and mutant IFRD1 protein structures were examined for their conformational flexibility and stability in order to evaluate the consequences of the point mutations. Molecular dynamics simulations were employed to assess the changes structural and stability of the protein. The impacts of the mutations were determined through the assessment of factors like changes in the secondary structure, Rg, and overall compression. These findings provide the valuable insights into the structural impact of the nsSNPs, thus lead to a better understanding of their possible functional implications.

4.15.1. Root Mean Square Deviation (RMSD)

To investigate the RMSD values for both wild-type and mutant proteins to the mutation impact on the protein structure. The wild-type IFRD1 protein showed constant simulation stability which was maintaining an RMSD having a range of 11.8-16.0 A throughout the entire simulation, indicating stable protein. On the other hand, the mutant IFRD1 protein containing both SNPs reached equilibrium at 20 nanoseconds after that reached stability only at 40nn followed by a period of stability at 80 ns constant was again stable with RMSD values RMSD values ranging from 12-14 A. The mutant IFRD1 showed different behaviors in its overall L-RMSD for 20 nanoseconds time period with fluctuations, as compared impact of the target protein on the stability. Notably, our findings suggest that, compared to wild-type IFRD1, the

Substituted variants mutant protein. protein exhibit limited impact on significant structural differences. However, the mutant demonstrated structural changes as the simulation progressed, our results indicating altered protein dynamics and potential function as shown in the **Figure 4.14**. This, in turn, may significantly influence its behavior, interactions, and overall biology, potentially contributing to the pathogenesis of SCA18.

4.15.2. Root Mean Square Fluctuations (RMSF)

In the analysis of the RMSF fluctuations for each residue was also conducted to investigate the dynamic behavior of the protein. Our results indicate that the IFRD1 protein with mutant protein exhibited higher residue-level fluctuations compared to the wild-type IFRD1.Specifically, the RMSF values for the wild-type IFRD1 averaged around 12 Å, 4.0 Å, and 6.0 Å for residues between positions 50–200 and 300, respectively .In contrast, The mutant protein exhibited significantly higher RMSF values of 13.5 Å, 4.5 Å, 6.0 Å, and 12.0 Å at the particular regions .The highest peak values, observed at residue positions 150, 250, and 300, suggest a concerning increase in flexibility in the mutant protein, potentially destabilizing the protein structure and disrupting its functional dynamics, which may lead to a loss of specificity, reduced efficacy, or even protein aggregation as shown **Figure 4.15.**

4.15.3. Secondary structure elements (SSE)

Secondary Structure Elements (SSE) refer to the fundamental structural components of a protein, including α-helices, β-sheets, and loops (or coils)**.** These elements play a crucial role in maintaining the protein's overall architecture and function. Changes in SSE due to mutations can significantly impact protein stability, flexibility, and interactions, potentially altering its biological activity. To further understand protein properties and behavior, we performed a secondary structure analysis of wild-type IFRD1. Our finding suggests that the effects of the single nucleotide polymorphisms (SNPs) in our results showed that either the individual protein or the mutant protein mutation are responsible for alteration in the whole helix of secondary structural elements The results of the analysis in the form of percentages were as follows: 58.09 % A (wild-type), 56.70 % B. The graphs **Figure 4.16** showed that the IFRD1 model with mutant SNPs substitution model exhibited a significant increase in the secondary structure content being the wild-type and the mutant SNPs IFRD1

model.

4.15.4. Radius of gyration (Rg)

Rg is the Radius of Gyration, which is a very important scientific measure for identifying a protein's compactness and stability. This parameter specifies the root mean square distance of all atom regions from the protein's center of mass of the protein us indicating the related to the protein structure, flexibility and folding dynamics. The mutations could lead to changes in the radius of gyration which might suggest the changes in protein stability, association potential, or unfolding characteristics, which all of which are important for understanding the functional implications of structural modifications. The radius of gyration (Rg) compactness investigation helps to explore the structural dynamics of single-molecule wild-type and mutant IFRD1 proteins throughout the 100 ns molecular dynamics simulation. Rg distinguishes the protein’s compactness, which is important for stability and functionality. The wild-type protein showed stable Rg values at between 24.5 and 26.5 nm, which indicates that the protein consistent compact and stable structure during the whole simulation. Conversely, the mutant protein showed the highly fluctuations in Rg values, between the 24.0 and 26.5 nm. The mutant's average Rg is similar to that of the wild type. However, the greater variation suggests that the mutant protein is less compact and more flexible to the wild type as illustrated in the **Figure. 4.17.** This implies that the mutation induces subtle alterations in the protein's structure, potentially impacting its stability and interactions with other molecules.

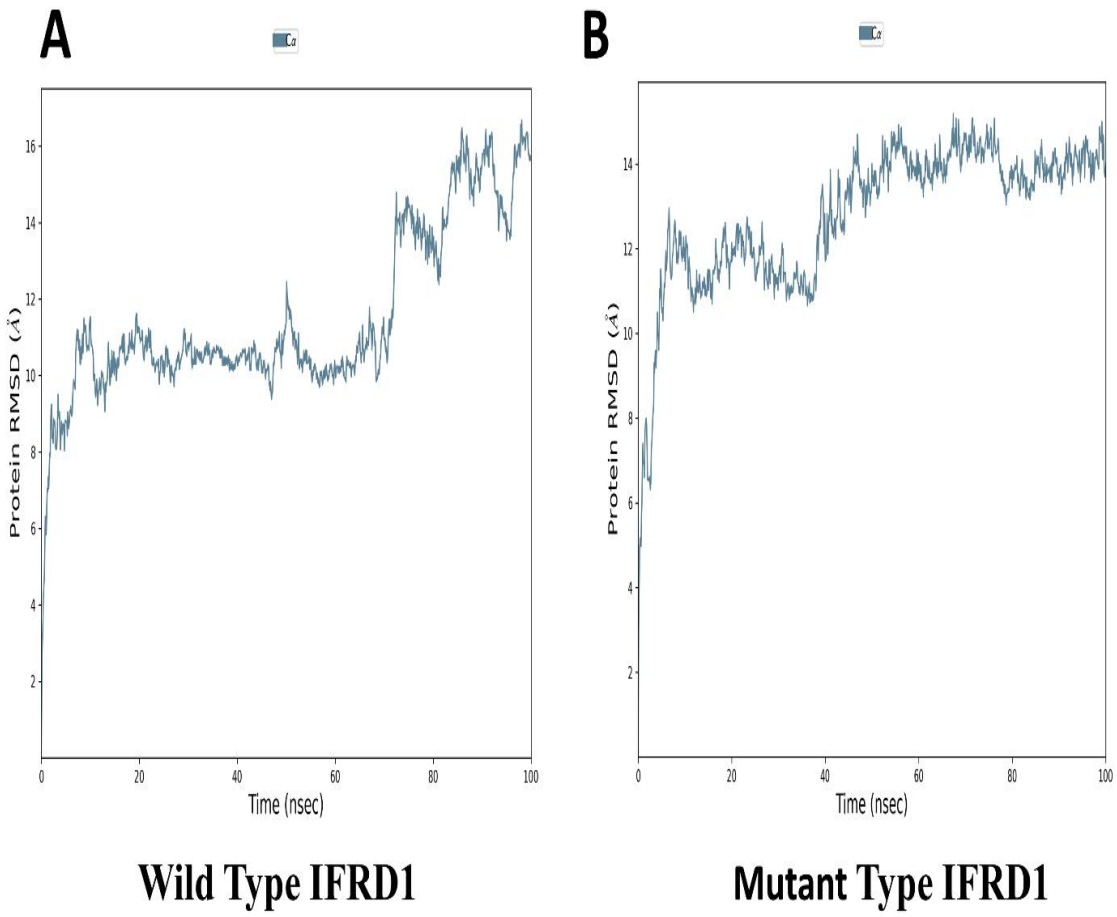


Figure 4.14. The RMSD analysis of wild-type and mutant IFRD1 proteins.

The RMSD analysis of IFRD1 protein variants: (A) Wild-type maintains stable RMSD (11.8-16.0 Å) variants stabilize at 20 ns, (B) Mutant type exhibits stable RMSD (-12-14Å). Variants show structural differences, especially D179G, impacting protein dynamics and potentially function.

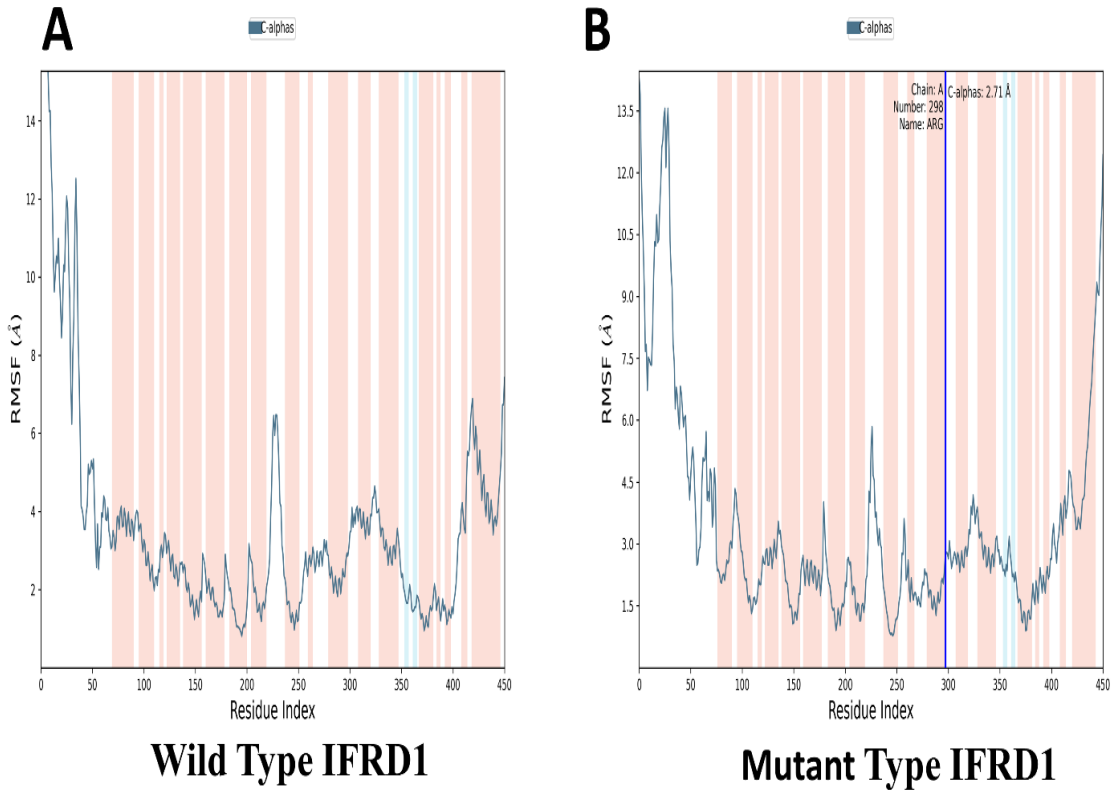


Figure 4. 15.RMSF analysis of the Wild Type and Mutant Type IFRD1 proteins

The RMSF analysis of IFRD1 protein variants: (A) Wild-type shows average RMSF of 12 Å, 4.0 Å, and 6.0Å; (B) Mutant type IFRD1 exhibit higher RMSF (13.5 Å, 4.5 Å, 6.0 Å, and 12.0 Å) at critical residues (50, 200, and 300), indicating increased flexibility and potential structural destabilization.

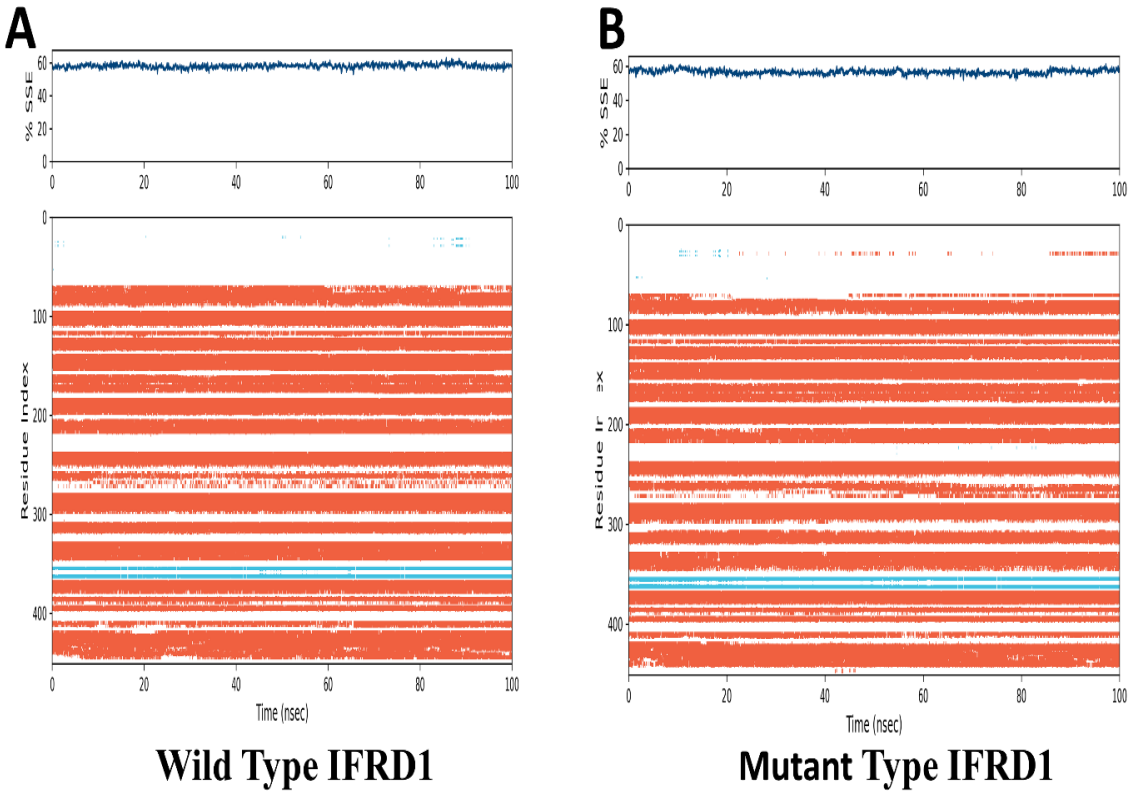


Figure 4. 16.Secondary structure analysis of wild-type and mutant type IFRD1

The Secondary structure analysis of wild-type and SNP-substituted IFRD1 proteins. (A) Wild-type IFRD1 shows 58.09 % helical content. B) Mutant type IFRD1 result in 56.70 % content. These findings indicate modifications in the helical elements due to SNPs, rather than affecting loops.

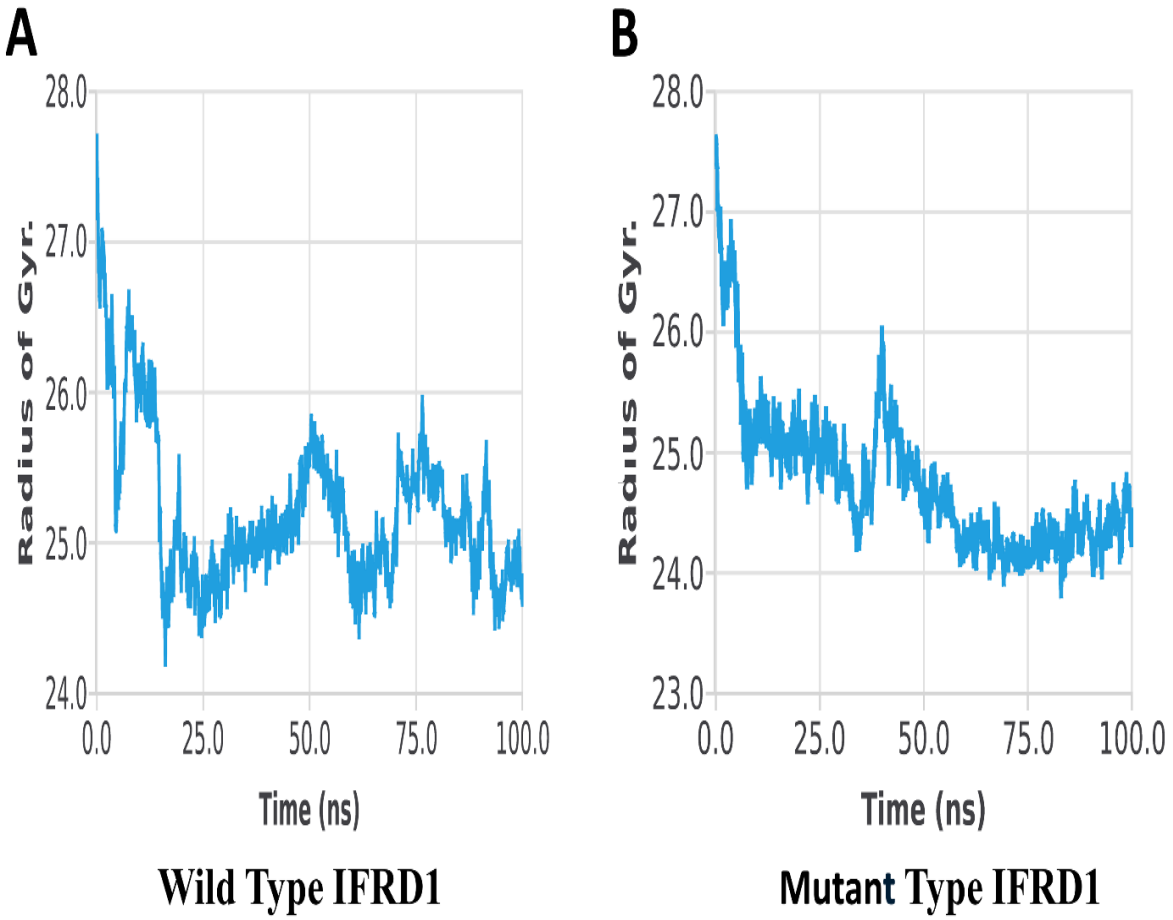


Figure 4. 17. Analysis of the radius of gyration for Wild-Type IFRD1 and Mutant IFRD1

The Comparison of Radius of Gyration in wild-type and SNP-substituted IFRD1 proteins. (A) Wild-type IFRD1 displayed 24.0 nm. (B) The mutant type IFRD1 showed 26.5 nm indicating increased flexibility and reduced stability and compactness.

Chapter: 05

Discussion

5.Discussion

Spinocerebellar ataxia 18 (SCA18) is a rare, autosomal dominant disorder classified under hereditary spinocerebellar ataxias [143]. These gradually progressive neurodegenerative Conditions lead to the gradual degeneration of the cerebellum and different parts of the central nervous system [144]. Mutations causing cerebellar atrophy can damage different brain regions, with effects that often vary even among members of the same family [17].
Mutations in the occur IFRD1 gene can lead to various diseases, including SCA18. Mutations in protein amino acids caused by nsSNPs can significantly affect the structure and function of proteins. NsSNPs cause most genetic disorders. Amino acid abnormalities lead to alterations in functionally important areas of proteins, which frequently come under significant selective pressure. The identification of specific amino acids and the differentiation between benign and pathogenic SNPs are significant challenges in experimental approaches that substantially influence disease progression [145].

The Insilco technique have the potential to classify the SNPs into three groups: benign, neutral, and harmful through the integration of several algorithms and SNPs information that are found in biological databases.[146]. The analysis of the amino acids formed due to mutations along with phylogenetic and structural data provides relatively accurate results [12]. A single Amino acid substitution mutation could change the function of the protein and increase the susceptibility to the disease. Moreover, single nucleotide polymorphisms have an impact on the enzyme activity and one of the modifications being structural instability of the proteins and disrupt both inter and protein hydrogen bonds as well as different hydrophobic interactions.

Current investigations in molecular biology and genetics of populations to find these SNPs-associated protein. Recently, the cutting-edge research in population genetics and molecular biology has identifying nonsynonymous single nucleotide polymorphisms (SNPs) that have a functional associated with proteins [147]. The in-silico methods used for identifying the location of single genes, prediction of gene transcripts, and analyze protein structures within cells [148].

The application of machine learning to genomics has not only influenced the process of discovering genetic markers through single nucleotide polymorphisms (SNPs) but has also

made a significant change in the way researchers mainly in the case of genetic markers associated with various genetic diseases. These studies have employed advanced computational methods, including a machine learning-based SNP-set analysis, to link genetic regions associated with specific diseases [50]. By utilizing sophisticated computational techniques, precisely identified the loci associated with diseases associated loci with remarkable precision [149].

Machine learning methods indicate efficiency in predicting complicated illnesses such as rheumatoid arthritis through SNP-based feature selection. The objective of our study was to identify and characterize functionally significant SNPs within the IFRD1 locus associated with SCA18 [150].We employed computational approaches to investigate 368 SNPs in the IFRD1 gene and discovered four nonsynonymous SNPs (nsSNPs)—D179G, R441Q, D308V, and G287V—that consistently indicated harmful consequences across various in-silico prediction tools. The analysis utilized predefined thresholds from tools including SIFT, PolyPhen-2, Meta-SNP, SNPs&GO, SNAP2, and PANTHER to assess the potential functional impacts of the identified variations on the IFRD1 protein. The results indicate that these nsSNPs could impair protein function, potentially disrupting cellular processes and contributing to the development of ataxia[151]. A machine learning approach was applied to examine disease-associated SNPs from a dataset of 8,872 variants using Random Forest, Support Vector Machine (SVM), and k-Nearest Neighbors (KNN). These models were employed to identify essential SNPs associated with specific diseases by analyzing their patterns and predictive relevance. The integration of these approaches provided a comprehensive framework for elucidating the genetic variants that influence disease susceptibility.

Furthermore, we used computational tools such as I-Mutant 2.0 and Mupro to analyze protein stability and to evaluate the structural effects of nsSNPs on the IFRD1 protein [151]. Our results indicated that machine learning predictions and functional analysis demonstrated that the nsSNPs (D179G, R441Q, D331G, and R396Q) significantly decreased protein stability (Table 4.4). This decrease in stability may lead to defective protein folding, misfolding, aggregation, and ultimately, a loss of function, potentially contributing to disease pathology. These results indicate that the identified nsSNPs could negatively affect the structure and stability of the IFRD1 protein, potentially contributing to the susceptibility to Spinocerebellar Ataxia 18. Conservation analysis through ConSurf tools [152] revealed that the residues

affected by the harmful nsSNPs (R441, D179, D331, and R396) are highly conserved regions emphasizing their functional importance. In the analysis, the Domain Conserved Site was identified as a critical region that was affected by these nsSNPs. The SOPMA server predicted that secondary structure analysis indicates mutations D179G and D331G in a β-sheet region and R441Q and R396Q in a coil region, leading to structural alterations that may compromise the integrity and function of IFRD1. These SNP-induced modifications play a critical role in destabilizing IFRD1, further supporting the prediction that the identified variants significantly impact its structure and functionality.

The finding provides significant insights into understanding to the evolutionary importance of these residues and their possible impact on the protein and susceptibility to Spinocerebellar Ataxia 18 disease. To analyzed the structural consequences of the identified nsSNPs was identified We applied both AlphaFold and SWISS-MODEL to predict the structures for the wild-type and mutant-IFRD1 proteins. Aspartic acid plays a significant role in various neurodegenerative diseases leading to the involvement of various mechanisms. Arginine is the primary source of nitric oxide (NO), which is a signaling molecule that plays a major role in neurodegenerative diseases. Aspartic acid plays the role of an excitatory neurotransmitter, Aspartic acid plays the role of an excitatory neurotransmitter, and is involved in neuronal death in diseases through NMDA receptor activation, a major contributor to excitotoxicity. Arginine is the main substrate for nitric oxide (NO), a signaling molecule significantly involved in the process of neurodegenerative disorders [154]. The Analysis by Project HOPE identifed mutations on the D179G, D1331G, R441Q, and R396Q are situated within important functional region of the IFRD1 protein, potentially disrupting its stability and function. In particular, the specific mutations affecting amino mutations at positions 179 and 1331 (D to G in a β-sheet region) and at positions 441 and 396 (R to Q in a coil region) could lead to a decrease in the protein’s functionality. These changes of the structural alterations caused by the mutations indicate that the mutations might be harmful.

Our MD simulations analysis demonstrates that the SNP variants have a major influence on the protein's structural stability, as the consequences significantly impact on the protein secondary structure content, less flexibility and also change in the protein's function. These findings confirm the previous research that indicated nsSNPs can impact the protein's

function by changing the structure's stability and dynamics [152],[155]. The increased secondary structure content in the SNPs substitution IFRD1 model may lead to reduced flexibility and limited impact on the protein function, potentially contributing to the pathogenesis of SCA18[151]. Furthermore, the decreased stability and SNPs of the IFRD1 protein may affect on the protein function and structure [156]. Our results suggest that the SNP possibly could have a significantly limited impact on the structure and the function of the IFRD1 protein, which might be the way that it is eventually contributing to the SCA18 disease progression.

Chapter: 06

Conclusion

6.Conclusion

In our extensive in-silico analysis and MD simulations, demonstrated that nonsynonymous single nucleotide (D179G and R441Q, R396Q and D331G) in the IFRD1 gene are majorly responsible for the gene's alteration significant effects on the protein structure and function, potentially contributing to the development of spinocerebellar ataxia type 18.Furthermore the machine learning method we employed displayed encouraging results, and feature selection improved precision model of the Random Forest 64%, Support Vector Machine (SVM) also 64%, while the alternative models even got to 67%.The models are thus confirmed to be very effective in predicting the impact of nsSNPs on disease outcomes. The identification of such as nsSNPs reveals possible therapeutic targets in the IFRD1 pathway, which subsequently lead to the exploration of new intervention strategies. The research investigation has a substantial impact on the comprehension of the genetic foundation of SCA18 as it reveals the involvement of non-synonymous SNPs in the control of the IFRD1 gene function. The genetic variations identified by our research have the potential contributes to improve the diagnostic, prognostic, and therapeutic strategies for SCA18 patients.

References

[1] N. E. Morton, J. M. Lalouel, J. F. Jackson, R. D. Currier, and S. Yee, 'Linkage studies in spinocerebellar ataxia (SCA)', *Am J Med Genet*, vol. 6, no. 3, pp. 251–257, 1980, doi: 10.1002/ajmg.1320060309.

[2] M. Rossi *et al.*, 'Autosomal dominant cerebellar ataxias: A systematic review of clinical features', *Eur J Neurol*, vol. 21, no. 4, pp. 607–615, 2014, doi: 10.1111/ene.12350.

[3] I. Miyai *et al.*, 'Cerebellar ataxia rehabilitation trial in degenerative cerebellar diseases', *Neurorehabil Neural Repair*, vol. 26, no. 5, pp. 515–522, 2012, doi: 10.1177/1545968311425918.

[4] A. H. Koeppen, 'The pathogenesis of spinocerebellar ataxia', *Cerebellum*, vol. 4, no. 1, pp. 62–73, 2005, doi: 10.1080/14734220510007950.

[5] R. Sullivan, W. Y. Yau, E. O'Connor, and H. Houlden, 'Spinocerebellar ataxia: an update', *J Neurol*, vol. 266, no. 2, pp. 533–544, 2019, doi: 10.1007/s00415-018-9076-4.

[6] U. Müller, 'Spinocerebellar ataxias (SCAs) caused by common mutations', *Neurogenetics*, pp. 235–250, 2021, doi: 10.1007/s10048-021-00662-5.

[7] H. L. Paulson, '基因的改变NIH Public Access', *Mol Cell Biochem*, vol. 23, no. 1, pp. 1–7, 2012, doi: 10.1097/WNO0b013e3181b416de.The.

[8] K. S. Bhalsing, V. Sowmya, M. Netravathi, S. Jain, and P. K. Pal, 'Spinocerebellar Ataxia (SCA) type 2 presenting with chorea', *Parkinsonism Relat Disord*, vol. 19, no. 12, pp. 1171–1172, 2013, doi: 10.1016/j.parkreldis.2013.08.004.

[9] H. Jacobi, T. Schaprian, T. Schmitz-Hübsch, M. Schmid, and T. Klockgether, 'Disease progression of spinocerebellar ataxia types 1, 2, 3 and 6 before and after ataxia onset', *Ann Clin Transl Neurol*, vol. 10, no. 10, pp. 1833–1843, 2023, doi: 10.1002/acn3.51875.

[10] M. Synofzik and W. Ilg, 'Motor training in degenerative spinocerebellar disease: Ataxia-specific improvements by intensive physiotherapy and exergames', *Biomed Res Int*, vol. 2014, 2014, doi: 10.1155/2014/583507.

[11] S. Jayadev and T. D. Bird, 'Hereditary ataxias: Overview', *Genetics in Medicine*, vol. 15, no. 9, pp. 673–683, 2013, doi: 10.1038/gim.2013.28.

[12] P. K. Panda, I. K. Sharawat, and L. Dawman, 'GRID2 Mutation-Related Spinocerebellar Ataxia Type 18: A New Report and Literature Review', *J Pediatr Genet*, vol. 11, no. 02, pp.

099–109, 2022, doi: 10.1055/s-0040-1721084.

[13] L. B. Hills *et al.*, 'Deletions in GRID2 lead to a recessive syndrome of cerebellar ataxia and tonic upgaze in humans', *Neurology*, vol. 81, no. 16, pp. 1378–1386, 2013, doi: 10.1212/WNL.0b013e3182a841a3.

[14] G. P. Rédei, 'Spinocerebellar Ataxia (SCA)', *Encyclopedia of Genetics, Genomics, Proteomics and Informatics*, pp. 1859–1859, 2008, doi: 10.1007/978-1-4020-6754-9_15990.

[15] J. Chataway *et al.*, 'Evidence that allelic variants of the spinocerebellar ataxia type 2 gene influence susceptibility to multiple sclerosis', pp. 91–96, 1999, doi: 10.007/s100489800059.

[16] M. Kim *et al.*, 'Late Diagnosis of Wilson Disease , Initially Presenting as Cerebellar Atrophy Mimicking Spinocerebellar Ataxia , by Multigene Panel Testing', pp. 500–503, 2020.

[17] R. Sullivan, W. Y. Yau, E. O'Connor, and H. Houlden, 'Spinocerebellar ataxia: an update', *J Neurol*, vol. 266, no. 2, pp. 533–544, 2019, doi: 10.1007/s00415-018-9076-4.

[18] S. Fujioka, C. Sundal, and Z. K. Wszolek, 'Autosomal dominant cerebellar ataxia type III: a review of the phenotypic and genotypic characteristics.', *Orphanet J Rare Dis*, vol. 8, pp. 1–13, 2013, doi: 10.1186/1750-1172-8-14.

[19] A. C. Ceylan, E. Acar Arslan, H. B. Erdem, H. Kavus, M. Arslan, and H. Topaloğlu, 'Autosomal recessive spinocerebellar ataxia 18 caused by homozygous exon 14 duplication in GRID2 and review of the literature', *Acta Neurol Belg*, vol. 121, no. 6, pp. 1457–1462, 2021, doi: 10.1007/s13760-020-01328-z.

[20] K. N. McFarland *et al.*, 'Repeat interruptions in spinocerebellar ataxia type 10 expansions are strongly associated with epileptic seizures', *Neurogenetics*, vol. 15, no. 1, pp. 59–64, 2014, doi: 10.1007/s10048-013-0385-6.

[21] Z. Brkanac *et al.*, 'Autosomal dominant sensory/motor neuropathy with ataxia (SMNA): Linkage to chromosome 7q22-q32', *American Journal of Medical Genetics - Neuropsychiatric Genetics*, vol. 114, no. 4, pp. 450–457, 2002, doi: 10.1002/ajmg.10361.

[22] A. C. Ceylan, E. Acar Arslan, H. B. Erdem, H. Kavus, M. Arslan, and H. Topaloğlu, 'Autosomal recessive spinocerebellar ataxia 18 caused by homozygous exon 14 duplication in GRID2 and review of the literature', *Acta Neurol Belg*, vol. 121, no. 6, pp. 1457–1462, 2021, doi: 10.1007/s13760-020-01328-z.

[23] S. Fujioka, C. Sundal, and Z. K. Wszolek, 'Autosomal dominant cerebellar ataxia type III: a review of the phenotypic and genotypic characteristics.', *Orphanet J Rare Dis*, vol. 8, pp. 1–

13, 2013, doi: 10.1186/1750-1172-8-14.

[24] E. K. Tan, 'Autosomal Dominant Spinocerebellar Ataxias: An Asian Perspective', *Canadian Journal of Neurological Sciences*, vol. 30, no. 4, pp. 361–367, 2003, doi: 10.1017/S0317167100003085.

[25] P. Lin, D. Zhang, G. Xu, and C. Yan, 'Identification of IFRD1 variant in a Han Chinese family with autosomal dominant hereditary spastic paraplegia associated with peripheral neuropathy and ataxia', *J Hum Genet*, vol. 63, no. 4, pp. 521–524, 2018, doi: 10.1038/s10038-017-0394-7.

[26] Z. Brkanac *et al.*, 'IFRD1 Is a Candidate Gene for SMNA on Chromosome 7q22-q23', *Am J Hum Genet*, vol. 84, no. 5, pp. 692–697, 2009, doi: 10.1016/j.ajhg.2009.04.008.

[27] U. Müller, 'Spinocerebellar ataxias (SCAs) caused by common mutations', *Neurogenetics*, pp. 235–250, 2021, doi: 10.1007/s10048-021-00662-5.

[28] C. Zhao, S. Datta, P. Mandal, S. Xu, and T. Hamilton, 'Stress-sensitive regulation of IFRD1 mRNA decay is mediated by an upstream open reading frame', *Journal of Biological Chemistry*, vol. 285, no. 12, pp. 8552–8562, 2010, doi: 10.1074/jbc.M109.070920.

[29] G. Park *et al.*, 'The transcriptional modulator Ifrd1 controls PGC-1α expression under short-term adrenergic stimulation in brown adipocytes', *FEBS Journal*, vol. 284, no. 5, pp. 784–795, 2017, doi: 10.1111/febs.14019.

[30] Y. Huang *et al.*, 'IFRD1 promotes tumor cells "low-cost" survival under glutamine starvation via inhibiting histone H1.0 nucleophagy', *Cell Discov*, vol. 10, no. 1, pp. 1–21, 2024, doi: 10.1038/s41421-024-00668-x.

[31] J. van Gaalen, P. Giunti, and B. P. Van de Warrenburg, 'Movement disorders in spinocerebellar ataxias', *Movement Disorders*, vol. 26, no. 5, pp. 792–800, 2011, doi: 10.1002/mds.23584.

[32] T. Adachi, M. Kitayama, T. Nakano, Y. Adachi, S. Kato, and K. Nakashima, 'Autopsy case of spinocerebellar ataxia type 31 with severe dementia at the terminal stage', *Neuropathology*, vol. 35, no. 3, pp. 273–279, 2015, doi: 10.1111/neup.12184.

[33] K. Seidel, S. Siswanto, E. R. P. Brunt, W. Den Dunnen, H. W. Korf, and U. Rüb, 'Brain pathology of spinocerebellar ataxias', *Acta Neuropathol*, vol. 124, no. 1, pp. 1–21, 2012, doi: 10.1007/s00401-012-1000-x.

[34] B. Trost, L. O. Loureiro, and S. W. Scherer, 'Discovery of genomic variation across a generation', *Hum Mol Genet*, vol. 30, no. R2, pp. R174–R186, 2021, doi:

10.1093/hmg/ddab209.

[35] H. Wealth Oyarieme and J. Otiti, ‘A Review on Genetic Variations within and between Populations: A Population Genetic Perspective’, *Article in American Research Journal of Biosciences*, vol. 9, no. 1, pp. 1–10, 2024, [Online]. Available: https://www.researchgate.net/publication/382869650

[36] E. E. Eichler, ‘Genetic Variation, Comparative Genomics, and the Diagnosis of Disease’, *New England Journal of Medicine*, vol. 381, no. 1, pp. 64–74, 2019, doi: 10.1056/nejmra1809315.

[37] C. S. Ku, E. Y. Loy, A. Salim, Y. Pawitan, and K. S. Chia, ‘The discovery of human genetic variations and their use as disease markers: Past, present and future’, *J Hum Genet*, vol. 55, no. 7, pp. 403–415, 2010, doi: 10.1038/jhg.2010.55.

[38] L. B. Jorde and S. P. Wooding, ‘Genetic Variation and Human Evolution’, *Nature Publishing Group*, vol. Volume 36, p. Pp.28-33, 2004.

[39] S. Azizzadeh-Roodpish, M. H. Garzon, and S. Mainali, ‘Classifying single nucleotide polymorphisms in humans’, *Molecular Genetics and Genomics*, vol. 296, no. 5, pp. 1161–1173, 2021, doi: 10.1007/s00438-021-01805-x.

[40] E. Vallejos-Vidal *et al.*, ‘Single-Nucleotide Polymorphisms (SNP) Mining and Their Effect on the Tridimensional Protein Structure Prediction in a Set of Immunity-Related Expressed Sequence Tags (EST) in Atlantic Salmon (Salmo salar)’, *Front Genet*, vol. 10, no. February, pp. 1–18, 2020, doi: 10.3389/fgene.2019.01406.

[41] A. Nusrath and B. Raiza P.T, ‘Review on Single Nucleotide Polymorphism Analysis Methods’, *Internaional Journal of Engineering Research & Technology (IJERT)*, vol. 3, no. 30, pp. 1–4, 2015, [Online]. Available: http://pmd.ddbj.nig.ac.jp

[42] L. N. Rodden *et al.*, ‘A non-synonymous single nucleotide polymorphism in SIRT6 predicts neurological severity in Friedreich ataxia’, *Front Mol Biosci*, vol. 9, no. September, pp. 1–13, 2022, doi: 10.3389/fmolb.2022.933788.

[43] M. J. Koretsky *et al.*, ‘Genetic risk factor clustering within and across neurodegenerative diseases’, *Brain*, vol. 146, no. 11, pp. 4486–4494, 2023, doi: 10.1093/brain/awad161.

[44] H. Padh, ‘Sequencing and comparative genome analysis of three Indians’, *Mammalian Genome*, vol. 32, no. 5, pp. 401–412, 2021, doi: 10.1007/s00335-021-09882-4.

[45] A. Ajith and U. Subbiah, ‘in silico screening of non-synonymous SNPs in human TUFT1 gene’, *Journal of Genetic Engineering and Biotechnology*, vol. 21, no. 1, p. 95, 2023, doi:

10.1186/s43141-023-00551-4.

[46] W. Zheng, C. Zhang, Y. Li, R. Pearce, E. W. Bell, and Y. Zhang, 'Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations', *Cell Reports Methods*, vol. 1, no. 3, p. 100014, 2021, doi: 10.1016/j.crmeth.2021.100014.

[47] M. Yirgu, M. Kebede, T. Feyissa, B. Lakew, A. B. Woldeyohannes, and M. Fikere, 'Single nucleotide polymorphism (SNP) markers for genetic diversity and population structure study in Ethiopian barley (Hordeum vulgare L.) germplasm', *BMC Genom Data*, vol. 24, no. 1, pp. 1–13, 2023, doi: 10.1186/s12863-023-01109-6.

[48] B. Dabhi and K. N. Mistry, 'in silico analysis of single nucleotide polymorphism (SNP) in human TNF-α gene', *Meta Gene*, vol. 2, pp. 586–595, 2014, doi: 10.1016/j.mgene.2014.07.005.

[49] D. S. W. Ho, W. Schierding, M. Wake, R. Saffery, and J. O'Sullivan, 'Machine learning SNP based prediction for precision medicine', *Front Genet*, vol. 10, no. MAR, pp. 1–10, 2019, doi: 10.3389/fgene.2019.00267.

[50] P. P. Silva *et al.*, 'A machine learning-based SNP-set analysis approach for identifying disease-associated susceptibility loci', *Sci Rep*, vol. 12, no. 1, pp. 1–10, 2022, doi: 10.1038/s41598-022-19708-1.

[51] M. Kang, S. Kim, D. Bin Lee, C. Hong, and K. B. Hwang, 'Gene - specific machine learning for pathogenicity prediction of rare BRCA1 and BRCA2 missense variants', *Sci Rep*, pp. 1–12, 2023, doi: 10.1038/s41598-023-37698-6.

[52] M. Khandakji *et al.*, 'BRCA1 -speci fi c machine learning model predicts variant pathogenicity with high accuracy', no. May 2023, pp. 315–323, 2025, doi: 10.1152/physiolgenomics.00033.2023.

[53] J. Gaudillo *et al.*, 'Machine learning approach to single nucleotide polymorphism-based asthma prediction', *PLoS One*, vol. 14, no. 12, pp. 1–12, 2019, doi: 10.1371/journal.pone.0225574.

[54] A. J. W. Lim *et al.*, 'Robust SNP-based prediction of rheumatoid arthritis through machine-learning-optimized polygenic risk score', *J Transl Med*, vol. 21, no. 1, pp. 1–17, 2023, doi: 10.1186/s12967-023-03939-5.

[55] E. G. Wakayu, 'Machine Learning Analysis of Single Nucleotide Polymorphism (SNP) Data to Predict Bone Mineral Density in African American Women', 2021.

[56] K. L. M. L. Hetzelt *et al.*, 'A case of severe autosomal recessive spinocerebellar ataxia type 18

with a novel nonsense variant in GRID2', *Eur J Med Genet*, vol. 63, no. 9, 2020, doi: 10.1016/j.ejmg.2020.103998.

[57] P. K. Panda, I. K. Sharawat, and L. Dawman, 'GRID2 Mutation-Related Spinocerebellar Ataxia Type 18: A New Report and Literature Review', *J Pediatr Genet*, vol. 11, no. 02, pp. 099–109, 2022, doi: 10.1055/s-0040-1721084.

[58] M. U. Manto, 'The wide spectrum of spinocerebellar ataxias (SCAs)', *Cerebellum*, vol. 4, no. 1, pp. 2–6, 2005, doi: 10.1080/14734220510007914.

[59] A. Baldan *et al.*, 'IFRD1 gene polymorphisms are associated with nasal polyposis in cystic fibrosis patients', *Rhinology journal*, vol. 53, no. 4, pp. 359–364, 2015, doi: 10.4193/rhin14.229.

[60] R. Xu, C. Peng, S. Xiao, and W. Zhuang, 'IFRD1 polymorphisms and gastric cancer risk in a Chinese population', *Medical Oncology*, vol. 31, no. 9, pp. 1–4, 2014, doi: 10.1007/s12032-014-0135-0.

[61] Y. Gu *et al.*, 'Altered Neutrophil Function', vol. 458, no. 7241, pp. 1039–1042, 2010, doi: 10.1038/nature07811.IFRD1.

[62] M. Manaz, Ö. F. Karasakal, E. Özkan Oktay, and M. Karahan, 'in silico analysis of missense SNPs in GABRA1, GABRB1, and GABRB3 genes associated with some diseases in neurodevelopmental disorders', *Egyptian Journal of Medical Human Genetics*, vol. 24, no. 1, 2023, doi: 10.1186/s43042-023-00446-6.

[63] M. S. Hossain, A. S. Roy, and M. S. Islam, 'in silico analysis predicting effects of deleterious SNPs of human RASSF5 gene on its structure and functions', *Sci Rep*, vol. 10, no. 1, pp. 1–14, 2020, doi: 10.1038/s41598-020-71457-1.

[64] S. U. Ahmad *et al.*, 'Computational screening and analysis of deleterious nsSNPs in human p14ARF (CDKN2A gene) protein using molecular dynamic simulation approach', *J Biomol Struct Dyn*, vol. 41, no. 9, pp. 3964–3975, 2023, doi: 10.1080/07391102.2022.2059570.

[65] M. Irfan, T. Iqbal, S. Hashmi, U. Ghani, and A. Bhatti, 'Insilico prediction and functional analysis of nonsynonymous SNPs in human CTLA4 gene', *Sci Rep*, vol. 12, no. 1, pp. 1–11, 2022, doi: 10.1038/s41598-022-24699-0.

[66] S. C. Das, M. A. Rahman, and S. Das Gupta, 'In-silico analysis unravels the structural and functional consequences of non-synonymous SNPs in the human IL-10 gene', *Egyptian Journal of Medical Human Genetics*, vol. 23, no. 1, 2022, doi: 10.1186/s43042-022-00223-x.

[67] M. Xu *et al.*, 'Genome Wide Association Study to predict severe asthma exacerbations in children using random forests classifiers', *BMC Med Genet*, vol. 12, no. 1, p. 90, 2011, doi: 10.1186/1471-2350-12-90.

[68] I. Joshi *et al., Artificial intelligence, big data and machine learning approaches in genome-wide SNP-based prediction for precision medicine and drug discovery*. INC, 2022. doi: 10.1016/B978-0-323-85713-0.00021-9.

[69] J. Cheng, A. Randall, and P. Baldi, 'Prediction of protein stability changes for single-site mutations using support vector machines', *Proteins: Structure, Function and Genetics*, vol. 62, no. 4, pp. 1125–1132, 2006, doi: 10.1002/prot.20810.

[70] S. T. Sherry *et al.*, 'DbSNP: The NCBI database of genetic variation', *Nucleic Acids Res*, vol. 29, no. 1, pp. 308–311, 2001, doi: 10.1093/nar/29.1.308.

[71] L. Phan *et al.*, 'The evolution of dbSNP: 25 years of impact in genomic research', *Nucleic Acids Res*, vol. 53, no. D1, pp. D925–D931, 2025, doi: 10.1093/nar/gkae977.

[72] K. L. Howe *et al.*, 'Ensembl 2021', *Nucleic Acids Res*, vol. 49, no. D1, pp. D884–D891, 2021, doi: 10.1093/nar/gkaa942.

[73] P. W. Harrison *et al.*, 'Ensembl 2024', *Nucleic Acids Res*, vol. 52, no. D1, pp. D891–D899, 2024, doi: 10.1093/nar/gkad1049.

[74] D. Cheng, C. Knox, N. Young, P. Stothard, S. Damaraju, and D. S. Wishart, 'PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites.', *Nucleic Acids Res*, vol. 36, no. Web Server issue, pp. 399–405, 2008, doi: 10.1093/nar/gkn296.

[75] J. Piñero *et al.*, 'DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants', *Nucleic Acids Res*, vol. 45, no. D1, pp. D833–D839, 2017, doi: 10.1093/nar/gkw943.

[76] C. L. Schoch *et al.*, 'NCBI Taxonomy: A comprehensive update on curation, resources and tools', *Database*, vol. 2020, no. 2, pp. 1–21, 2020, doi: 10.1093/database/baaa062.

[77] N. L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, and P. C. Ng, 'SIFT web server: Predicting effects of amino acid substitutions on proteins', *Nucleic Acids Res*, vol. 40, no. W1, pp. 452–457, 2012, doi: 10.1093/nar/gks539.

[78] M. Y. Behairy, Al. A. Abdelrahman, H. Y. Abdallah, E. E. D. A. Ibrahim, A. A. Sayed, and M. M. Azab, 'in silico analysis of missense variants of the C1qA gene related to infection and

autoimmune diseases', *J Taibah Univ Med Sci*, vol. 17, no. 6, pp. 1074–1082, 2022, doi: 10.1016/j.jtumed.2022.04.014.

[79] A. Ali *et al.*, 'in silico Tools for Analysis of Single-Nucleotide Polymorphisms in the Bovine Transferrin Gene', *Animals*, vol. 12, no. 6, pp. 1–16, 2022, doi: 10.3390/ani12060693.

[80] I. Adzhubei, D. M. Jordan, and S. R. Sunyaev, *Predicting functional effect of human missense mutations using PolyPhen-2*, no. SUPPL.76. 2013. doi: 10.1002/0471142905.hg0720s76.

[81] K. Shinwari *et al.*, 'Novel high-risk missense mutations identification in FAT4 gene causing Hennekam syndrome and Van Maldergem syndrome 2 through molecular dynamics simulation', *Inform Med Unlocked*, vol. 37, no. November 2022, p. 101160, 2023, doi: 10.1016/j.imu.2023.101160.

[82] I. A. Adzhubei *et al.*, 'A method and server for predicting damaging missense mutations', *Nat Methods*, vol. 7, no. 4, pp. 248–249, 2010, doi: 10.1038/nmeth0410-248.

[83] E. Capriotti, R. Calabrese, P. Fariselli, P. L. Martelli, R. B. Altman, and R. Casadio, 'WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation.', *BMC Genomics*, vol. 14 Suppl 3, no. Suppl 3, 2013, doi: 10.1186/1471-2164-14-s3-s6.

[84] H. Mi, N. Guo, A. Kejariwal, and P. D. Thomas, 'PANTHER version 6: Protein sequence and function evolution data with expanded representation of biological pathways', *Nucleic Acids Res*, vol. 35, no. SUPPL. 1, pp. 247–252, 2007, doi: 10.1093/nar/gkl869.

[85] H. Mi, A. Muruganujan, J. T. Casagrande, and P. D. Thomas, 'Large-scale gene function analysis with the panther classification system', *Nat Protoc*, vol. 8, no. 8, pp. 1551–1566, 2013, doi: 10.1038/nprot.2013.092.

[86] H. Mi and P. Thomas, 'PANTHER pathway: an ontology-based pathway database coupled with data analysis tools.', *Methods Mol Biol*, vol. 563, pp. 123–140, 2009, doi: 10.1007/978-1-60761-175-2_7.

[87] E. Capriotti, R. B. Altman, and Y. Bromberg, 'Collective judgment predicts disease-associated single nucleotide variants.', *BMC Genomics*, vol. 14 Suppl 3, no. Suppl 3, 2013, doi: 10.1186/1471-2164-14-s3-s2.

[88] T. Yasmin, 'in silico comprehensive analysis of coding and non-coding SNPs in human mTOR protein', *PLoS One*, vol. 17, no. 7 July, pp. 1–23, 2022, doi: 10.1371/journal.pone.0270919.

[89] M. Hecht, Y. Bromberg, and B. Rost, 'Better prediction of functional effects for sequence

variants From VarI-SIG 2014: Identification and annotation of genetic variants in the context of structure, function and disease', *BMC Genomics*, vol. 16, no. 8, pp. 1–12, 2015.

[90] X. Yu and S. Sun, 'Comparing a few SNP calling algorithms using low-coverage sequencing data', *BMC Bioinformatics*, vol. 14, no. 1, 2013, doi: 10.1186/1471-2105-14-274.

[91] B. Greshake, P. E. Bayer, H. Rausch, and J. Reda, 'openSNP-A crowdsourced web resource for personal genomics', *PLoS One*, vol. 9, no. 3, pp. 1–9, 2014, doi: 10.1371/journal.pone.0089204.

[92] M. Hajiloo *et al.*, 'Breast cancer prediction using genome wide single nucleotide polymorphism data', *BMC Bioinformatics*, vol. 14, no. SUPPL13, 2013, doi: 10.1186/1471-2105-14-S13-S3.

[93] H. Soueidan and M. Nikolski, 'Machine learning for metagenomics: methods and tools', *Metagenomics*, vol. 1, no. 1, pp. 1–19, 2016, doi: 10.1515/metgen-2016-0001.

[94] P. Larrañaga *et al.*, 'Machine learning in bioinformatics', *Brief Bioinform*, vol. 7, no. 1, pp. 86–112, 2006, doi: 10.1093/bib/bbk007.

[95] K. Maharana, S. Mondal, and B. Nemade, 'A review: Data pre-processing and data augmentation techniques', *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, 2022, doi: 10.1016/j.gltp.2022.04.020.

[96] E. Ibrahimi *et al.*, 'Overview of data preprocessing for machine learning applications in human microbiome research', *Front Microbiol*, vol. 14, no. October, pp. 1–8, 2023, doi: 10.3389/fmicb.2023.1250909.

[97] A. Amato and V. Di Lecce, 'Data preprocessing impact on machine learning algorithm performance', *Open Computer Science*, vol. 13, no. 1, 2023, doi: 10.1515/comp-2022-0278.

[98] K. M. Kahloot and P. Ekler, 'Algorithmic Splitting: A Method for Dataset Preparation', *IEEE Access*, vol. 9, pp. 125229–125237, 2021, doi: 10.1109/ACCESS.2021.3110745.

[99] J. Pebrianto, A. Ahmad, and M. Isnan, 'ScienceDirect ScienceDirect Machine Learning Approach for Single Nucleotide Polymorphism Selection in Genetic Testing Results', *Procedia Comput Sci*, vol. 227, pp. 46–54, 2023, doi: 10.1016/j.procs.2023.10.501.

[100] S. K. Saha, S. Sarkar, and P. Mitra, 'Feature selection techniques for maximum entropy based biomedical named entity recognition', *J Biomed Inform*, vol. 42, no. 5, pp. 905–911, 2009, doi: 10.1016/j.jbi.2008.12.012.

[101] T. K. Ho, 'Random decision forests', *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, vol. 1, pp. 278–282, 1995, doi:

10.1109/ICDAR.1995.598994.

[102] Z. Jin, J. Shang, Q. Zhu, C. Ling, W. Xie, and B. Qiang, 'RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12343 LNCS, pp. 503–515, 2020, doi: 10.1007/978-3-030-62008-0_35.

[103] M. W. Libbrecht and W. S. Noble, 'Machine learning in genetics and genomics', *Nature Review Genetics*, vol. 16, no. 6, pp. 321–332, 2015, doi: 10.1038/nrg3920.Machine.

[104] B. Petersen, T. N. Petersen, P. Andersen, M. Nielsen, and C. Lundegaard, 'A generic method for assignment of reliability scores applied to solvent accessibility predictions', *BMC Struct Biol*, vol. 9, pp. 1–10, 2009, doi: 10.1186/1472-6807-9-51.

[105] S. Huang, C. A. I. Nianguang, P. Penzuti Pacheco, S. Narandes, Y. Wang, and X. U. Wayne, 'Applications of support vector machine (SVM) learning in cancer genomics', *Cancer Genomics Proteomics*, vol. 15, no. 1, pp. 41–51, 2018, doi: 10.21873/cgp.20063.

[106] D. Coomans and D. L. Massart, 'Alternative k-nearest neighbour rules in supervised pattern recognition. Part 1. k-Nearest neighbour classification by using alternative voting rules', *Anal Chim Acta*, vol. 136, no. C, pp. 15–27, 1982, doi: 10.1016/S0003-2670(01)95359-0.

[107] J. Listgarten *et al.*, 'Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms', *Clinical Cancer Research*, vol. 10, no. 8, pp. 2725–2737, 2004, doi: 10.1158/1078-0432.CCR-1115-03.

[108] E. Capriotti, P. Fariselli, and R. Casadio, 'I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure', *Nucleic Acids Res*, vol. 33, no. SUPPL. 2, pp. 306–310, 2005, doi: 10.1093/nar/gki375.

[109] E. Capriotti and R. B. Altman, 'A new disease-specific machine learning approach for the prediction of cancer-causing missense variants', *Genomics*, vol. 98, no. 4, pp. 310–317, 2011, doi: 10.1016/j.ygeno.2011.06.010.

[110] E. Capriotti, P. Fariselli, I. Rossi, and R. Casadio, 'A three-state prediction of single point mutations on protein stability changes', *BMC Bioinformatics*, vol. 9, no. SUPPL. 2, pp. 1–9, 2008, doi: 10.1186/1471-2105-9-S2-S6.

[111] E. Capriotti and R. B. Altman, 'Improving the prediction of disease-related variants using protein three-dimensional structure', *BMC Bioinformatics*, vol. 12, no. SUPPL. 4, 2011, doi: 10.1186/1471-2105-12-S4-S3.

[112] H. Ashkenazy *et al.*, 'ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules', *Nucleic Acids Res*, vol. 44, no. W1, pp. W344–W350, 2016, doi: 10.1093/NAR/GKW408.

[113] H. Ashkenazy, E. Erez, E. Martz, T. Pupko, and N. Ben-Tal, 'ConSurf 2010: Calculating evolutionary conservation in sequence and structure of proteins and nucleic acids', *Nucleic Acids Res*, vol. 38, no. SUPPL. 2, pp. 529–533, 2010, doi: 10.1093/nar/gkq399.

[114] P. A. Barnard, 'Secondary school structure, organisational learning capacity and learning organisations: a systemic contribution', *International Journal of Educational Management*, vol. 34, no. 8, pp. 1253–1264, 2020, doi: 10.1108/IJEM-01-2020-0037.

[115] G. Deléage, 'ALIGNSEC: viewing protein secondary structure predictions within large multiple sequence alignments', *Bioinformatics*, vol. 33, no. 24, pp. 3991–3992, 2017, doi: 10.1093/bioinformatics/btx521.

[116] M. S. Klausen *et al.*, 'NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning', *Proteins: Structure, Function and Bioinformatics*, vol. 87, no. 6, pp. 520–527, 2019, doi: 10.1002/prot.25674.

[117] M. H. Høie *et al.*, 'NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning', *Nucleic Acids Res*, vol. 50, no. W1, pp. W510–W515, 2022, doi: 10.1093/nar/gkac439.

[118] M. Dorn, M. B. E Silva, L. S. Buriol, and L. C. Lamb, 'Three-dimensional protein structure prediction: Methods and computational strategies', *Comput Biol Chem*, vol. 53, no. PB, pp. 251–276, 2014, doi: 10.1016/j.compbiolchem.2014.10.001.

[119] J. Jumper *et al.*, 'Highly accurate protein structure prediction with AlphaFold', *Nature*, vol. 596, no. 7873, pp. 583–589, 2021, doi: 10.1038/s41586-021-03819-2.

[120] O. Kovalevskiy, J. Mateos-Garcia, and K. Tunyasuvunakool, 'AlphaFold two years on: Validation and impact', *Proc Natl Acad Sci U S A*, vol. 121, no. 34, pp. 1–6, 2024, doi: 10.1073/pnas.2315002121.

[121] A. Waterhouse *et al.*, 'SWISS-MODEL: Homology modelling of protein structures and complexes', *Nucleic Acids Res*, vol. 46, no. W1, pp. W296–W303, 2018, doi: 10.1093/nar/gky427.

[122] T. Schwede, J. Kopp, N. Guex, and M. C. Peitsch, 'SWISS-MODEL: An automated protein homology-modeling server', *Nucleic Acids Res*, vol. 31, no. 13, pp. 3381–3385, 2003, doi:

10.1093/nar/gkg520.

[123] J. Xu, F. Jiao, and L. Yu, 'Protein structure prediction using threading', *Methods in Molecular Biology*, vol. 413, pp. 91–121, 2007, doi: 10.1385/1-59745-574-1:91.

[124] S. E. Adeniji, S. Uba, and A. Uzairu, 'in silico study for evaluating the binding mode and interaction of 1, 2, 4-triazole and its derivatives as potent inhibitors against Lipoate protein B (LipB)', *J King Saud Univ Sci*, vol. 32, no. 1, pp. 475–485, 2020, doi: 10.1016/j.jksus.2018.07.014.

[125] A. Corporate, 'Discovery Studio Life Science Modeling and Simulations', *Researchgate.Net*, pp. 1–8, 2008.

[126] S. W. Lim *et al.*, 'Functional and structural analysis of non-synonymous single nucleotide polymorphisms (nsSNPs) in the MYB oncoproteins associated with human cancer', *Sci Rep*, vol. 11, no. 1, pp. 1–14, 2021, doi: 10.1038/s41598-021-03624-x.

[127] H. Venselaar, T. A. H. te Beek, R. K. P. Kuipers, M. L. Hekkelman, and G. Vriend, 'Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces', *BMC Bioinformatics*, vol. 11, no. 1, p. 548, 2010, doi: 10.1186/1471-2105-11-548.

[128] M. I. Mustafa, N. S. Murshed, A. H. Abdelmoneim, and A. M. Makhawi, 'in silico analysis of the functional and structural consequences of SNPs in human ARX gene associated with EIEE1', *Inform Med Unlocked*, vol. 21, p. 100447, 2020, doi: 10.1016/j.imu.2020.100447.

[129] A. Han, H. J. Kang, Y. Cho, S. Lee, Y. J. Kim, and S. Gong, 'SNPΔomain: A web resource of single nucleotide polymorphisms (SNPs) within protein domain structures and sequences', *Nucleic Acids Res*, vol. 34, no. WEB. SERV. ISS., pp. 642–644, 2006, doi: 10.1093/nar/gkl323.

[130] A. Marchler-Bauer *et al.*, 'CDD: A Conserved Domain Database for the functional annotation of proteins', *Nucleic Acids Res*, vol. 39, no. SUPPL. 1, pp. 225–229, 2011, doi: 10.1093/nar/gkq1189.

[131] A. Marchler-Bauer *et al.*, 'CDD: A conserved domain database for interactive domain family analysis', *Nucleic Acids Res*, vol. 35, no. SUPPL. 1, pp. 237–240, 2007, doi: 10.1093/nar/gkl951.

[132] S. Hunter *et al.*, 'InterPro: The integrative protein signature database', *Nucleic Acids Res*, vol. 37, no. SUPPL. 1, pp. 211–215, 2009, doi: 10.1093/nar/gkn785.

[133] T. Paysan-Lafosse *et al.*, 'InterPro in 2022', *Nucleic Acids Res*, vol. 51, no. D1, pp. D418–

D427, 2023, doi: 10.1093/nar/gkac993.

[134] M. P. E. Kevin J. Bowers, Edmond Chow, Huafeng Xu, Ron O. Dror, F. D. S. Brent A. Gregersen, John L. Klepeis, Istvan Kolossvary, Mark A. Moraes, D. E. S. John K. Salmon, Yibing Shan, and D. E. Shaw, 'Microsoft Word - sc06-paper-aug01-edmond.doc', *Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters*, pp. 1–13, 2006.

[135] K. J. Bowers *et al.*, 'Scalable algorithms for molecular dynamics simulations on commodity clusters', *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing, SC'06*, 2006, doi: 10.1145/1188455.1188544.

[136] S. V. G. Reddy, K. T. Reddy, V. V. Kumari, and S. H. Basha, 'Molecular docking and dynamic simulation studies evidenced plausible immunotherapeutic anticancer property by Withaferin A targeting indoleamine 2,3-dioxygenase', *J Biomol Struct Dyn*, vol. 33, no. 12, pp. 2695–2709, 2015, doi: 10.1080/07391102.2015.1004834.

[137] D. Shivakumar, J. Williams, Y. Wu, W. Damm, J. Shelley, and W. Sherman, 'Prediction of absolute solvation free energies using molecular dynamics free energy perturbation and the opls force field', *J Chem Theory Comput*, vol. 6, no. 5, pp. 1509–1519, 2010, doi: 10.1021/ct900587b.

[138] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, 'Comparison of simple potential functions for simulating liquid water', *J Chem Phys*, vol. 79, no. 2, pp. 926–935, 1983, doi: 10.1063/1.445869.

[139] L. G. Ferreira, R. N. Dos Santos, G. Oliva, and A. D. Andricopulo, *Molecular docking and structure-based drug design strategies*, vol. 20, no. 7. 2015. doi: 10.3390/molecules200713384.

[140] J. Cheng, Y. Liu, and Y. Ma, 'Protein secondary structure prediction based on integration of CNN and LSTM model', *J Vis Commun Image Represent*, vol. 71, p. 102844, 2020, doi: 10.1016/j.jvcir.2020.102844.

[141] L. Zhao, J. Li, W. Zhan, X. Jiang, and B. Zhang, 'Prediction of protein secondary structure by the improved TCN-BiLSTM-MHA model with knowledge distillation', *Sci Rep*, vol. 14, no. 1, pp. 1–21, 2024, doi: 10.1038/s41598-024-67403-0.

[142] S. Ren, 'Effects of arginine in therapeutic protein formulations: a decade review and perspectives', *Antib Ther*, vol. 6, no. 4, pp. 265–276, 2023, doi: 10.1093/abt/tbad022.

[143] H. L. Paulson, '基因的改变NIH Public Access', *Mol Cell Biochem*, vol. 23, no. 1, pp. 1–7,

2012, doi: 10.1097/WNO0b013e3181b416de.The.

[144] F. Ghorbani *et al.*, 'Copy Number Variant Analysis of Spinocerebellar Ataxia Genes in a Cohort of Dutch Patients with Cerebellar Ataxia', *Neurol Genet*, vol. 9, no. 1, pp. 1–8, 2023, doi: 10.1212/NXG.0000000000200050.

[145] R. A. Hubner and R. S. Houlston, 'Single nucleotide polymorphisms and cancer susceptibility', *The Molecular Basis of Human Cancer*, vol. 8, no. 66, pp. 231–239, 2016, doi: 10.1007/978-1-59745-458-2_14.

[146] S. M. Wakil *et al.*, 'A genome-wide association study reveals susceptibility loci for myocardial infarction/coronary artery disease in Saudi Arabs', *Atherosclerosis*, vol. 245, pp. 62–70, 2016, doi: 10.1016/j.atherosclerosis.2015.11.019.

[147] C. Yanchus *et al.*, 'A noncoding single-nucleotide polymorphism at 8q24 drives IDH1-mutant glioma formation', *Science (1979)*, vol. 378, no. 6615, pp. 68–78, 2022, doi: 10.1126/science.abj2890.

[148] B. K. Ghazi *et al.*, 'in silico Structural and Functional Analyses of NLRP3 Inflammasomes to Provide Insights for Treating Neurodegenerative Diseases', *Biomed Res Int*, vol. 2023, 2023, doi: 10.1155/2023/9819005.

[149] E. Widen, T. G. Raben, L. Lello, and S. D. H. Hsu, 'Machine Learning Prediction of Biomarkers from SNPs and of Disease Risk from Biomarkers in the UK Biobank', 2021.

[150] F. Ghadiri, 'A machine-learning approach for nonalcoholic steatohepatitis susceptibility estimation', vol. 41, no. October, pp. 475–482, 2022.

[151] S. Saxena *et al.*, 'In-silico analysis of deleterious single nucleotide polymorphisms of PNMT gene', *Mol Simul*, vol. 48, no. 16, pp. 1411–1425, 2022, doi: 10.1080/08927022.2022.2094922.

[152] E. Özkan Oktay, T. Kaman, Ö. F. Karasakal, and V. Enisoğlu Atalay, 'in silico Prediction and Molecular Docking of SNPs in NRP1 Gene Associated with SARS-COV-2', *Biochem Genet*, vol. 1, no. 0123456789, 2023, doi: 10.1007/s10528-023-10409-6.

[153] M. Holeček, 'Aspartic Acid in Health and Disease', *Nutrients*, vol. 15, no. 18, 2023, doi: 10.3390/nu15184023.

[154] X. L. Liu, S. Sato, W. Dai, and N. Yamanaka, 'The protective effect of hepatocyte growth-promoting factor (pHGF) against hydrogen peroxide-induced acute lung injury in rats', *Medical Electron Microscopy*, vol. 34, no. 2, pp. 92–102, 2001, doi: 10.1007/s007950170003.

[155] S. Falahi, A. G. Karaji, F. Koohyanizadeh, A. Rezaiemanesh, and F. Salari, 'A comprehensive

in silico analysis of the functional and structural impact of single nucleotide polymorphisms (SNPs) in the human IL-33 gene', *Comput Biol Chem*, vol. 94, no. June, p. 107560, 2021, doi: 10.1016/j.compbiolchem.2021.107560.

[156] L. N. Rodden *et al.*, 'A non-synonymous single nucleotide polymorphism in SIRT6 predicts neurological severity in Friedreich ataxia', *Front Mol Biosci*, vol. 9, no. September, pp. 1–13, 2022, doi: 10.3389/fmolb.2022.933788.