ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

VISUALIZATION BASED ANALYSIS OF GENE NETWORKS USING HIGH DIMENSIONAL MODEL REPRESENTATION

M.Sc. THESIS

Pınar GÜLER

Department of Computational Science and Engineering

Computational Science and Engineering Programme

JULY 2024

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

VISUALIZATION BASED ANALYSIS OF GENE NETWORKS USING HIGH DIMENSIONAL MODEL REPRESENTATION

M.Sc. THESIS

Pınar GÜLER
(702211009)

Department of Computational Science and Engineering

Computational Science and Engineering Programme

Thesis Advisor: Asst. Prof. Dr. Süha TUNA

JULY 2024

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

YÜKSEK BOYUTLU MODEL GÖSTERİLİM KULLANILARAK GEN AĞLARININ GÖRSELLEŞTİRME TABANLI ANALİZİ

YÜKSEK LİSANS TEZİ

Pınar GÜLER
(702211009)

Hesaplamalı Bilim ve Mühendislik Anabilim Dalı

Hesaplamalı Bilim ve Mühendislik Programı

Tez Danışmanı: Dr. Öğr. Üyesi Süha TUNA

TEMMUZ 2024

Pınar GÜLER, a M.Sc. student of ITU Graduate School student ID 702211009 successfully defended the thesis entitled "VISUALIZATION BASED ANALYSIS OF GENE NETWORKS USING HIGH DIMENSIONAL MODEL REPRESENTATION", which she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Asst. Prof. Dr. Süha TUNA** Istanbul Technical University	.....
Jury Members :	**Assoc. Prof. Dr. Burcu TUNGA** Istanbul Technical University	.....
	Assoc. Prof. Dr. Yelda TARKAN ARGÜDEN Istanbul University - Cerrahpaşa	.....

Date of Submission : **24 May 2024**
Date of Defense : **1 July 2024**

To my spouse,

FOREWORD

Completing this master's thesis has been a rewarding journey for me, fostering both personal and professional growth. I extend my sincere gratitude to my advisor, Dr. Süha TUNA, for his unwavering support, guidance, and motivating encouragement throughout this academic endeavour. Additionally, I express deep gratitude to my spouse, my family and friends, whose presence and encouragement have been a constant source of strength and inspiration, sustaining me through every phase of this process.

July 2024 Pınar GÜLER

TABLE OF CONTENTS

	Page
FOREWORD	**ix**
TABLE OF CONTENTS	**xi**
ABBREVIATIONS	**xiii**
SYMBOLS	**xv**
LIST OF TABLES	**xvii**
LIST OF FIGURES	**xix**
SUMMARY	**xxi**
ÖZET	**xxiii**
1. INTRODUCTION	**1**
1.1 Purpose of the Thesis	3
1.2 Literature Review	6
2. METHODS	**11**
2.1 Chaos Game Representation	11
2.2 VARCH	14
2.3 High Dimensional Model Representation	19
2.4 Support Vector Machines	26
3. IMPLEMENTATION	**29**
3.1 mTOR Dataset	29
3.2 Numerical Experiments	30
3.2.1 Balanced dataset	30
3.2.2 Imbalanced dataset	31
4. RESULTS	**33**
5. CONCLUSION	**41**
REFERENCES	**43**
CURRICULUM VITAE	**49**

ABBREVIATIONS

DNA : Deoxyribonucleic Acid
RNA : Ribonucleic Acid
SNP : Single Nucleotide Polymorphism
GWAS : Genome-Wide Association Studies
CGR : Chaos Game Representation
HDMR : High-Dimensional Model Representation
SVM : Support Vector Machines
RBF : Radial Basis Function
BCE : Binary Cross-Entropy
PPI : Protein-Protein Interaction
ChIP : Chromatin Immunoprecipitation
PRS : Polygenic Risk Score
LD : Linkage Disequilibrium
1D-CNN : 1D Convolutional Neural Network
2D-CNN : 2D Convolutional Neural Network
VLC : Variant Logic Constructor
BioGRID : Biological General Repository for Interaction Datasets
ChIP : Chromatin ImmunoPrecipitation
RICTOR : Rapamycin-Insensitive Companion of mTOR
mTOR : Mammalian Target of Rapamycin
KEGG : Kyoto Encyclopedia of Genes and Genomes
NCBI : National Center for Biotechnology Information
mTORC1 : Mammalian Target of Rapamycin Complex 1
mTORC2 : Mammalian Target of Rapamycin Complex 2
STRING : Search Tool for Retrieval of Interacting Genes

SYMBOLS

A	**:** Adenine
T	**:** Thymine
C	**:** Cytosine
G	**:** Guanine
T_{α}^{i}	**:** The x-axis element in frequency matrix of VARCH
T_{β}^{h}	**:** The y-axis element in frequency matrix of VARCH
$\perp$	**:** Variant Logic Operator
$+$	**:** Variant Logic Operator
$-$	**:** Variant Logic Operator
$\top$	**:** Variant Logic Operator
$k-mer$	**:** Substrings of length k in a sequence
N	**:** The length of DNA Sequence in VARCH Method
Len	**:** The length of a subsequence in VARCH Method
m	**:** The number of subsequences in VARCH Method
H	**:** Principal HDMR Tensor
K	**:** Linear Kernel of SVM
$\mathbf{s}_i$	**:** HDMR support vectors
h_0	**:** 0-dimensional (scalar) HDMR component
$\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3$	**:** 1-dimensional (vector) HDMR components

LIST OF TABLES

Page

Table 2.1 : The relationships in variant logic construction. **14**
Table 2.2 : The relationship between subsets and values. **15**
Table 4.1 : VARCH results for varying *Len* and *m* parameters ($T_\alpha = 9$, $T_\beta = 6$).. **37**
Table 4.2 : Classification accuracy performances of T_α and T_β pairs ($Len = 20, m = 20$). **38**
Table 4.3 : VARCH ($Len = 10$, $m = 160$) and CGR ($n = 5$) results of 400 patient/control data. **38**
Table 4.4 : VARCH ($Len = 10, m = 160$) and CGR ($n = 5$) results of 100 patient/400 control data. **39**

LIST OF FIGURES

Page

Figure 2.1 : 32×32 CGR images of mTOR gene: GSK3B (top-left), RICTOR (top- right), SLC38A9 (bottom-left), MTOR (bottom-right). **12**
Figure 2.2 : CGR process of genomic sequence data. **13**
Figure 2.3 : Subsegmentation of the genomic sequence data. **16**
Figure 2.4 : VARCH images of varying T_α and T_β values ($Len = 10$, $m = 200$).. **18**
Figure 2.5 : Binary classification using SVM linear kernel. **27**
Figure 4.1 : Flowchart of the proposed Gene Network Analysis Process. **34**
Figure 4.2 : Classification accuracies depending on $k-mer$ parameter for CGR and HDMR (400 patient/control data). **35**
Figure 4.3 : Varying effects of VARCH's sequence length on accuracy, F1 score, recall and BCE loss. **36**

VISUALIZATION BASED ANALYSIS OF GENE NETWORKS USING HIGH DIMENSIONAL MODEL REPRESENTATION

SUMMARY

Genetic studies have revolutionized our understanding of the biological mechanisms underlying health and disease. By exploring the intricate details of the human genome, researchers can identify genetic variations that contribute to various phenotypic outcomes. One of the key advancements in this field is gene network analysis, which examines the complex interactions between genes and how they regulate cellular processes. This approach provides a comprehensive view of the biological systems and uncovers the pathways involved in disease mechanisms. Genome-Wide Association Studies (GWAS) play a pivotal role among the methodologies utilized in gene network analysis. GWAS involves scanning the genome for slight variations, known as single nucleotide polymorphisms (SNPs), that occur more frequently in individuals with a particular disease or trait than in those without. By identifying these associations, GWAS helps pinpoint genetic factors contributing to disease susceptibility and progression, paving the way for personalized medicine and targeted therapeutic strategies. By integrating various variant analysis techniques, researchers can develop a deeper understanding of the genetic architecture of diseases, leading to significant advancements in diagnostics, treatment, and prevention.

Gene network and pathway analyses are essential components of genetic studies, offering insights into genes' complex interactions and functions within a biological systems. However, both face significant computational challenges, mainly when dealing with high-dimensional genomic data. Analyzing vast datasets containing gene expression profiles and genetic variations demands sophisticated computational methods capable of handling their scale and complexity. Conventional statistical methods frequently require assistance to become effective, demanding complex computational approaches like data visualization, network modeling, and machine learning algorithms. In addition, the complexity of biological networks and pathways makes analysis even more complicated, necessitating the use of powerful computational tools to interpret regulatory mechanisms and simulate complex biological processes correctly. Overcoming these challenges is crucial for gaining deeper insights into gene networks and pathways, thereby advancing our understanding of their roles in health and disease. In pathway analysis, scientists employ data collected from many sources, such as Genome-Wide Association Studies (GWAS), to identify target genes and connect them to known pathways using Kyoto Encyclopedia of Genes and Genomes (KEGG) databases. However, pathway analysis presents major computing challenges, especially when large, high-dimensional genomic datasets are involved. Researchers have developed innovative methods such as High Dimensional Model Representation (HDMR), Chaos Game Representation (CGR), and visual

analysis of DNA sequences based on a variant logic construction method called VARCH to overcome these challenges. By mapping genetic sequences into visual representations, these innovative approaches can help identify potential genetic markers and better understand biological processes. These computational methods must be included in gene network and pathway investigations to fully understand the complex architecture of genetic interactions and how they affect health and diseases.

In this thesis, we harnessed three sophisticated computational methodologies: Chaos Game Representation, visual analysis of DNA sequences based on variant logic construction called VARCH, and High Dimensional Model Representation, each offering unique contributions to the variant analysis, respectively CGR, a prevalent technique in bioinformatics, translates genetic sequences into visually interpretable diagrams, clarifying complex structures and patterns in the sequences. On the other hand, VARCH converts sequences into a feature space, successfully capturing each aspect of their complexity and uncertainty. These techniques are effective instruments in our search for potential genetic markers that might help us distinguish between the patient and control groups in our investigation. Furthermore, we utilized HDMR for dimension reduction, an essential technique for simplifying the complex structure in high-dimensional genomic data. By condensing data dimensions, HDMR facilitated more efficient and accurate classification, enabling us to uncover sensitive genetic relationships and patterns that might have remained hidden otherwise. Integrating these computational techniques provided robust solutions for analyzing genetic data from the mTOR pathway, enriching our comprehension of the genetic mechanisms supporting various phenotypic outcomes.

In our study, we begin on a mission to deepen our comprehension of the intricate genetic patterns intertwined with diverse phenotypic outcomes. Focusing on genetic data sourced from the mTOR pathway, we leveraged state-of-the-art computational methodologies to unravel hidden insights. Our primary objective was to assess the efficacy of CGR, VARCH, and HDMR in gene network analyses. As we analyzed the data, the results were quite compelling. Both CGR and VARCH methods demonstrated notable accuracy in genetic classification, with VARCH exhibiting a significant edge over CGR in terms of accuracy and sensitivity metrics. This superiority was underscored by VARCH's ability to considerably minimize binary cross-entropy (BCE) loss values, demonstrating the ability to reduce errors in predictions. However, we examined the computing overheads associated with each methodology in detail, providing insight into the challenging trade-off between computational complexity and accuracy. Despite the more significant parameters, VARCH's computational requirements were apparent, although its performance was better than CGR's. Our study demonstrates the potential of computational tools for unraveling gene complexities while also acting as an essential reminder of how crucial it is to overcome the complex environment of computational constraints carefully, helping researchers search for the best possible method selection and optimization.

YÜKSEK BOYUTLU MODEL GÖSTERİLİM KULLANILARAK GEN AĞLARININ GÖRSELLEŞTİRME TABANLI ANALİZİ

ÖZET

Genetik çalışmalar, hastalıklarda altta yatan biyolojik mekanizmaların anlaşılmasında devrim niteliğinde ilerlemeler sağlamıştır. Bu ilerlemeler, genetik bilgilere dayalı sağlık hizmetlerinin kişiselleştirilmesinde önemli bir rol oynamaktadır. Araştırmacılar, insan genomunun karmaşık mekanizmalarını keşfederek, çeşitli fenotipik sonuçlara katkıda bulunan genetik varyasyonları tespit edebilirler. Bu keşifler, hastalıkların genetik temellerini anlamada ve bu sayede risk grubunda yer alan bireylerin erken tespitinde ve aynı zamanda yeni tedavi stratejileri geliştirmede kritik öneme sahiptir. Bu alandaki temel ilerlemelerden biri, gen ağ analizidir. Bu yaklaşım genler arasındaki karmaşık etkileşimleri ve hücresel süreçleri nasıl düzenlediklerini inceler. Gen ağ analizleri, genlerin birbirleriyle olan ilişkilerini ve bu ilişkilerin biyolojik fonksiyonlar üzerindeki etkilerini ortaya çıkartarak daha kapsamlı bir biyolojik anlayış sağlar ve bu sayede hastalık mekanizmalarında yer alan yolları açığa çıkarır. Genom çapında birleşim çalışmaları (GWAS), gen ağ analizinde kullanılan metodlar arasında kilit bir rol oynamaktadır. GWAS, büyük ölçekli veri setlerinden genetik ilişkileri belirlemek için kullanılan güçlü bir araçtır. GWAS, belirli bir hastalığı veya hastalığa ait genetik özelliğine sahip olan bireylerde daha sık görülen küçük varyasyonları, tek nükleotid polimorfizmleri (SNP'ler), genom çapında taramayı içermektedir. Bu ilişkileri ayırt ederek, GWAS hastalık duyarlılığını ve ilerlemesine katkıda bulunan genetik faktörleri belirler ve kişiselleştirilmiş tıp hizmetlerine ve hedefe yönelik tedaviler için temel oluşturur. Kişiselleştirilmiş tıp, bireylerin genetik profillerine dayalı olarak daha etkili tedavi yaklaşımları geliştirmeyi mümkün kılar. Farklı genetik analiz tekniklerinin bir araya getirilmesi, araştırmacıların hastalıkların genetik yapılarını daha iyi anlamalarını sağlar ve bu da teşhis, tedavi ve önleme alanlarında önemli ilerlemelere yol açar. Bu entegrasyon, genetik bilgilerin klinik uygulamalarda nasıl kullanılabileceğini belirler ve sağlık alanındaki yenilikçi çözümler için temel oluşturur.

Gen ağları analizi ve yol analizi, genetik çalışmaların vazgeçilmez bileşenleridir ve biyolojik sistemler içindeki genlerin karmaşık etkileşimlerini ve fonksiyonlarını açıklar. Bu yöntemler, genetik araştırmalarda derinlemesine bilgi sağlar ve hastalıkların moleküler temelini aydınlatmada büyük bir rol oynar. Ancak, her iki yöntemde de özellikle yüksek boyutlu genom verileriyle uğraşırken önemli hesaplama zorluklarıyla karşılaşılır. Bu zorluklar, verilerin karmaşıklığı ve veri setlerinin büyüklüğü nedeniyle ortaya çıkar. Bu sebepler nedeniyle veri analizi süreci oldukça karmaşık hale gelmektedir. Gen profilleri ve genetik varyasyonlar içeren geniş veri kümelerinin analizi, ölçeğini ve karmaşıklığını yönetebilen gelişmiş hesaplama yöntemlerini gerektirir. Geleneksel istatistiksel yöntemler yerine veri görselleştirme, ağ modelleme ve makine öğrenimi algoritmaları gibi karmaşık

hesaplama yaklaşımlarını gerektirir. Bu yöntemler, verilerin daha etkin bir şekilde analiz edilmesini ve yorumlanmasını sağlar.

Gen ağlarının karmaşıklığı ve gen regülasyon mekanizmalarını anlamak, karmaşık biyolojik süreçleri doğru bir şekilde simüle etmek için güçlü hesaplama araçları gerektirir; bu zorlukları aşmak, gen ağlarının ve yollarının sağlık ve hastalık üzerindeki rollerini daha derinlemesine anlamamıza olanak tanır. Özellikle kompleks hastalıkların incelenmesinde, gen ağlarının doğru bir şekilde modellenmesi ve analiz edilmesi kritik bir öneme sahiptir. Yol analizinde, araştırmacılar, GWAS gibi çeşitli kaynaklardan elde edilen verileri kullanarak hedef genleri belirler ve Kyoto Genlerin ve Genomların Ansiklopedisi (KEGG) veritabanlarını kullanarak bu genleri bilinen yollarla ilişkilendirirler. KEGG veritabanı, biyolojik yolları ve bunların genlerle olan ilişkilerini anlamada önemli bir kaynaktır. Ancak, özellikle yüksek boyutlu ve büyük genom veri kümeleriyle çalışıldığında, yol analizi çalışmaları önemli hesaplama zorluklarıyla karşılaşır. Bu zorlukların üstesinden gelmek için DNA dizilerinin görsel analizini içeren yenilikçi yöntemler geliştirilmiştir; bunlar arasında Kaos Oyunu Temsili (CGR) ve varyant mantık yapısına dayalı DNA dizilerinin görsel analiz yöntemi olan VARCH bulunmaktadır.

Genetik dizilerin görsel temsili, potansiyel genetik sinyal yolları belirlemeye ve biyolojik süreçleri daha iyi anlamaya yardımcı olur. Görsel analiz yöntemleri, karmaşık genetik verilerin anlaşılmasını kolaylaştırır ve genetik araştırmalarda yeni perspektifler sunar. Bu hesaplama yöntemlerinin gen ağları ve yol analizlerine entegre edilmesi, genetik etkileşimlerin karmaşık yapısını tam olarak anlamak ve bunların hastalıklar üzerindeki etkilerini anlamak için hayati öneme sahiptir. Böylece, genetik araştırmalarda daha hassas ve doğru sonuçlar elde edilir.

Tez kapsamında, gen ağları analizimize katkılar sunan iki adet görselleştirme ve bir data ayrıştırma yöntemi kullanıldı. Bu yöntemler, genetik verilerin daha iyi anlaşılmasını ve analiz edilmesini sağlayan güçlü araçlardır. Kaos Oyunu Temsili, varyant mantık yapısına dayalı DNA dizilerinin görsel analizi olan VARCH ve Yüksek Boyutlu Model Gösterilim (HDMR) yöntemlerini kullandık. CGR, biyoinformatikte yaygın kullanılan bir yöntem olup, genetik dizileri görsel olarak yorumlanabilir diyagramlara çevirerek dizilerdeki karmaşık yapıları ve desenleri açıklar. CGR yöntemi, genetik dizilerin görselleştirilmesi ile büyük veri setlerindeki yapısal özellikleri ortaya çıkarır. Öte yandan, VARCH, gen sekanslarını bir özellik vektörüne dönüştürerek, bu sekansların karmaşık yapılarını başarıyla yakalamaktadır. VARCH, genetik verilerin detaylı analizini mümkün kılarak, araştırmacıların gen sekanslarındaki önemli paternleri tespit etmelerini sağlar. Her iki yöntem de, araştırmamızda hasta ve kontrol gruplarını ayırt edebilecek potansiyel genetik nitelikleri aramamızda etkili olmuştur. Bu sayede, belirli hastalık fenotiplerine özgü genetik işaretçileri daha kolay belirleyebildik.

Bunun ile birlikte, genlerin karmaşık yapılarının basitleştirilmesi için kullanılabilecek önemli bir teknik olan boyut indirgeme için HDMR methodunu kullandık. HDMR, yüksek boyutlu veri setlerinde önemli bilgileri koruyarak veri boyutunu azaltmak için etkili bir yöntemdir. Kullanılan mTOR veri setindeki verilerin görselleştirme yöntemlerinden elde edilen N boyutlu tensörlere uygulanan HDMR, hesaplama açısından belirgin bir verimlilik sağlamış ve aynı zamanda yüksek sınıflandırma

performansı göstermiştir. Bu teknik, büyük veri setlerinde hızlı ve doğru analiz yapmayı mümkün kılarak araştırma sürecini hızlandırmıştır. Buna ek olarak, gizli kalan genetik ilişkileri ve desenlerin ortaya çıkarılmasına olanak tanımıştır. Bu sayede, genetik verilerin derinlemesine analizi ile hasta ve kontrol gruplarına ait verilerin yüksek doğruluk yüzdesiyle sınıflandırılmasına dair bulgular elde edilmiştir. Bu hesaplama tekniklerinin entegrasyonu, mTOR yolu genetik verilerinin analizinde tutarlı çözümler sunarak, çeşitli fenotipik sonuçları destekleyen genetik mekanizmaların daha iyi anlaşılmasını sağlamıştır. Sonuç olarak, bu yöntemler, genetik araştırmalarda daha derinlemesine ve kapsamlı analizler yapmamıza olanak tanıyarak, genetik varyasyonların hastalıklar üzerindeki etkilerini daha iyi kavramamızı sağlamıştır.

Tez kapsamında, çeşitli fenotipik sonuçlarla ilişkilendirilen karmaşık genetik desenlerin daha derinlemesine anlaşılmasına katkı sağlayacak ve yüksek sınıflandırma performansı sağlayacak hesaplama yöntemlerin incelenmesi amaçlanmıştır. Bu kapsamda, mTOR yolu ile ilişkili genetik verilerin analizi için güncel hesaplama teknikleri kullanılarak daha ayrıntılı sonuçlar elde edilmeye çalışılmıştır. mTOR yolu kaynaklı genetik verilere odaklanılarak, genlere ait nitelikleri açığa çıkarmak için güncel hesaplama metodlarını kullanılmıştır. Temel amaç, CGR, VARCH ve HDMR yöntemlerinin gen ağ analizlerindeki etkinliğini değerlendirmek olup bu yöntemlerin etkinliğini değerlendirirken, her birinin genetik verileri ne kadar yüksek oranda temsil edebildiği ve analiz sonuçlarının doğruluğunun incelenmesidir. Elde edilen genetik veriler sınıflandırma yöntemi ile analiz edildiğinde, önemli sonuçlar elde edildiği gözlemlenmiştir. Hem CGR hem de VARCH yöntemleri genetik sınıflamada dikkate değer bir doğruluk gösterdi. Bu doğruluk oranları, genetik varyasyonların doğru bir şekilde tespit edilmesi ve sınıflandırılması açısından önemli bir başarıdır. VARCH metodu hem dengeli dağılmış datasetinde hem de dengesiz (imbalanced) datasetlerle, doğruluk ve hassasiyet metrikleri açısından CGR'ye göre belirgin bir avantaj sağladı. VARCH yöntemi bu üstünlüğü, genetik verilerin daha hassas ve detaylı bir şekilde analiz edilmesine olanak tanımaktadır. Bu üstünlük, VARCH yöntemi hata oranlarını önemli ölçüde azaltma yeteneği ile daha da vurgulandı. Özellikle, genetik hastalıkların erken teşhisi ve tedavi stratejilerinin geliştirilmesinde VARCH yöntemi kritik bir rol oynayabilir. Bununla birlikte, her yöntemin ilişkilendirilen hesaplama maliyetlerini detaylı olarak inceledik ve hesaplama karmaşıklığı ile doğruluk arasındaki denge hakkında bir anlayış sağladık. Bu denge, yöntemlerin pratik uygulamalarda ne kadar verimli olabileceğini ve büyük veri setleriyle çalışırken hangi yöntemin daha uygun olacağını belirlememize yardımcı oldu. Optimize edilecek parametrelerin daha fazla ve performansı CGR'den daha iyi olmasına karşın VARCH'ın hesaplama gereksinimleri açık bir şekilde ortaya konmuştur. Bu çalışma, gen karmaşıklıklarını çözme konusunda hesaplama araçlarının potansiyelini ortaya koyarken, hesaplama kısıtlamalarını dikkatlice aşmanın önemini vurgulamakta ve araştırmacıların en uygun yöntemi seçme ve optimize etme konusunda rehberlik etmektedir.

1. INTRODUCTION

Human genetic data, including all of an individual's genetic information in DNA, are the key to understanding the underlying principles of heredity, genetic variation, and disease susceptibility. Human genome data is the entire collection of genetic information encoded in a human organism's DNA. This contains all the genes (coding sections) and non-coding DNA regions, such as regulatory sequences and repetitive patterns, that comprise the whole genome [1]. In conclusion, human genetic data focuses on particular variants or sequences within the genome, whereas human genome data includes an individual's or species' complete genetic blueprint. Multimodal data collection is a critical component in unraveling the complicated fabric of human biology, providing essential insights into the genetic drivers of many phenotypic characteristics, complex illnesses, and fundamental biological processes [2].

Human genetic and gene data studies are widely investigating the genetic composition of individuals or groups to understand the complexities of genetics and genomics and its implications for health and diseases. These extensive studies explore the depths of genetic data to identify distinct genetic variants, such as mutations or single nucleotide polymorphisms (SNPs), and their relationships to specific phenotypes, illnesses, or behaviors [3]. Complex genotyping techniques, high-throughput genomic sequencing, and intricate bioinformatics analyses are just a few sophisticated tools and techniques that researchers have used to carefully examine the unique structure, functionality, and diversity of the human genome.

Genome-Wide Association Studies (GWAS) are a prominent method to identify genetic variations linked to complicated disorders. The early 2000s saw the rise in popularity of GWAS. Since then, GWAS have gained popularity as a method for locating genetic variations linked to both standard and complicated disorders [4] [5]. GWAS examines all genetic differences in the genome to determine the potential relationship between these variations and the likelihood of acquiring

particular characteristics or diseases. In order to discover information about the genetic components of complex disorders, GWAS uses SNPs to locate and map genetic variations linked to various traits and diseases. For that purpose, researchers have used large databases of SNPs to investigate gene-gene and gene-environment interactions and evaluate associations between genetic variations and diseases. GWAS have proven helpful in locating shared genomic variations associated with different traits and illnesses [5] [6]. The knowledge gained from these initiatives also opens the door for creating and applying personalized medicine strategies that target customizing medical interventions and treatments to each patient's unique genetic profile. This enables healthcare delivery and enhance patient outcomes.

In genomics research, pathways analysis is performed to examine the connections and interactions between genes and the outputs they produce within biological pathways [7]. These pathways are sets of chemical reactions or signaling cascades that control several aspects of the cell, including metabolism, cell cycle, and the immune system. Genes are categorized in route analysis according to their biological roles or participation in particular pathways. This arrangement enables scientists to evaluate the collective activity of genes within a pathway and to comprehend the potential effects of genetic variants or changes in gene expression on biological processes. Pathway analysis frequently integrates omics data such as gene expression, SNP data, or protein-protein interactions to identify dysregulated pathways linked to diseases.

The growing number of GWAS has necessitated advances in computing techniques and processes used to analyze large amounts of data. This mutually beneficial development have been emerged necessary by the increasing intricacy of genetic research and the urgent need for increased analytical precision. For this reasons, advancements in computational techniques and computers have been a significant factor in the success of human genetic research. Computational developments have sped up the processing and interpreting of enormous volumes of genomic data, starting with early computational models and statistical methods and continuing with high-performance computing and machine learning techniques [8] [9] [10]. The development of genome browsers, data visualization platforms, and bioinformatic tools has facilitated the finding of new information in human genetics [11] [12]. Additionally, these

technologies have simplified data mining and integration. Furthermore, open-access data repositories and computational tools have expedited scientific discoveries and encouraged collaborative research, contributing to the globalization of genetic data access.

Initially, GWAS used traditional statistical techniques to determine the relationships between genetic variations and phenotypic characteristics. These approaches usually require the evaluation of hypotheses using methods such as logistic regression or chi-squared tests [13] [14]. However, as datasets became larger and more complicated, traditional statistical methods found it difficult to handle the entire volume of genetic data and account for confounding variables such as population segmentation and repeated testing modifications.

In GWAS, logistic regression with LASSO regularization is commonly used for SNP preselection and has proven effective in classifying especially Crohn's Disease patients [15]. This involves using penalized regression algorithms like LASSO (L1 penalty) and ridge regression (L2 penalty) to model phenotypes as a linear sum of genetic variants, with regularization in place to limit coefficient magnitude. Ensemble learning algorithms such as Random Forest play a valuable role in genetic data analysis by capturing complex genetic relationships and identifying significant variants [16] [17]. In addition, recent studies have highlighted the effectiveness of neural networks in GWAS. For instance, CNN architecture is used to predict phenotypes [18] [19], and dense neural networks are used to distinguish between patients and controls [15]. These methods show potential in identifying significant genetic markers and predicting disease phenotypes.

1.1 Purpose of the Thesis

Variant analyses are essential in biomedical research because they help us comprehend various biological processes and disorders. Researchers can get critical insights into the genetic foundation of illnesses and responsiveness to therapies by analyzing an individual's genetic composition. Furthermore, genome analysis makes it possible to identify genetic variants, mutations, and biomarkers linked to the development or risk of disease. Personalized medicine benefits greatly from this knowledge as it enables

medical professionals to customize interventions and therapies based on a patient's genetic profile. Additionally, variant analysis aids in the discovery of new therapeutic targets and the creation of focused medicines, both of which eventually enhance patient outcomes.

GWAS often face several challenges. Firstly, the significant computational costs, especially with multidimensional data, can be an obstacle. Genetic data analysis usually involves numerous variables, leading to computational complexities requiring substantial resources and time. Secondly, the high dimensionality of genetic datasets exacerbates these issues. The immense volume of data increases the computational burden and can obstruct efficient analysis. Additionally, these factors may affect classification accuracy rates in variant analyses. The complexity of the data and computational constraints can impair the effectiveness of classification algorithms. This can result in low-accurate classification into different groups, potentially leading to suboptimal performance and reduced classification accuracy. Addressing these issues is crucial for improving variant analysis techniques' effectiveness, accuracy, and scalability. Ultimately, this will enhance our understanding of genetic phenomena and their impacts on health and disease.

In this thesis, we study utilized two genetic visualization tools and a multidimensional data decomposition method called High Dimensional Model Representation to enable gene classification between patient and control groups. The aim was to understand better genetic patterns associated with different phenotypic outcomes. The visualization tools enabled us to identify potential biological indicators or genetic signatures that distinguish between the two groups by providing insights into the genome data's underlying structure. The multidimensional data decomposition technique helped manage the complexity of the genome datasets, making our classification study more accurate and efficient. The combined use of these techniques facilitated the examination and categorization of gene sequences, advancing the field of genetics and paving the way for personalized therapy and disease diagnosis.

As the proposed method, we used the Chaos Game Representation (CGR) as our visualization technique. This method allowed us to create CGR images by adjusting

the resolution according to the chosen k-mer value [20]. We aligned the CGR images of each individual within their respective patient and control groups, which resulted in two tensors, one for each group. These 3-D tensors served as the basis for further genetic data analysis, aiding in identifying trends and differences between the groups.

The second visualization technique we utilized is the VARCH method [21]. This method transformed the input gene sequences into VARCH images based on four parameters that control the representations' accuracy. We tested if these parameters could enhance accuracy compared to CGR. Similar to CGR, VARCH produced two tensors, each representing a group. These tensors were also used for further comparative analysis and for exploring genetic patterns related to different phenotypic outcomes.

In both methods, we used the High Dimensional Model Representation (HDMR) for data decorrelation, dimension reduction, and feature extraction tasks [22]. After applying HDMR, we transformed the individual images into three vectors and combined them into a single vector for each participant. This resulted in one feature vector for each sample in the patient and control groups. We then employed a Support Vector Machines (SVM) algorithm on these feature vectors to classify individuals into their groups based on extracted genetic data.

Finally, we analyzed the feature matrices using the SVM algorithm and compared the classification accuracy of the visualization methods. We also fine-tuned the method parameters to observe their impact on results. Besides classification accuracy, we considered other machine learning metrics such as F1 score, sensitivity, and binary cross-entropy (BCE) loss. We evaluated computational costs throughout these processes, aiming to assess the performance of the visualization methods in terms of classification accuracy and their efficacy in capturing relevant patterns and optimizing model performance.

The results from the analyses are discussed in detail in Section 4. This section also explores the implications of these findings and suggests potential areas for future research. Section 5 focuses on addressing the limitations of the proposed method and suggesting directions for further investigation, which involves refining the

methodologies used, exploring additional parameters or techniques, and broadening the scope of analysis to fill any remaining gaps or uncertainties. In highlighting these areas for improvement and future exploration, the study aims to contribute to ongoing advancements in the field and inspire further research.

1.2 Literature Review

Advancements in genomic research, mainly through GWAS, have significantly enhanced our understanding of genetic variations and their impact on human health and diseases. GWAS focuses on identifying genetic variations associated with diseases. These studies analyze genetic markers, especially SNPs, across the entire genome to uncover how these variations may influence disease risk [3] [4]. The methodology of GWAS involves comparing the genomes of individuals with a specific disease to those without, aiming to detect correlations between specific genetic variants and disease presence. This approach helps pinpoint specific genetic variances linked to diseases. To ensure the associations' reliability, GWAS requires analyzing many individuals.

One of the strengths of GWAS is its use of statistical methods to determine the importance of genetic associations. Furthermore, GWAS often includes diverse populations to capture the full spectrum of genetic variations across different ethnic groups, enhancing the generalizability of the findings [4]. The implications of GWAS are profound for precision medicine. Findings from these studies contribute to risk prediction, improve diagnostic accuracy, and prepare for personalized treatment strategies. By understanding the genetic basis of diseases, medical professionals can develop more targeted therapies tailored to an individual's genetic makeup, thereby increasing efficacy and reducing side effects.

Despite the transformative potential of GWAS, it has its own set of difficulties. A significant challenge is the need for large sample sizes to detect meaningful genetic associations, especially for complex diseases influenced by multiple genetic and environmental factors. Additionally, understanding the complexities of genetic interactions and their contributions to diseases remains challenging, as these interactions often involve multiple genes and pathways.

Also, to better understand the SNPs that are the subject of GWAS, SNPs are common genetic variations that occur when a single nucleotide (A, T, C, or G) at a specific position in the genome differs among individuals in a population. They are the most abundant form of genetic variation in the human genome and play a significant role in determining an individual's traits, susceptibility to diseases, and response to medications. The interrelationship between SNPs and pathway analysis is crucial for understanding the genetic basis of complex characteristics and disorders. Pathway analysis is a technique that elucidates how genetic variations impact biological processes. This is achieved by mapping SNPs to genes and subsequently performing gene-set enrichment analysis [3]. Considering the cumulative effects of many SNPs within pathways, this technique significantly improves statistical power and sheds light on the biological processes driving genetic relationships. Key points about SNPs include:

- Genetic Variation: SNPs represent variations in a single nucleotide within the DNA sequence.
- Biological Significance: SNPs can influence gene expression, protein function, and overall phenotype. They are associated with various traits, diseases, and drug responses.
- GWAS: SNPs are commonly used as genetic markers in GWAS to identify associations between specific genetic variants and diseases or traits.
- Population Genetics: SNPs can vary in frequency across different populations, providing insights into genetic diversity and evolutionary relationships.
- Clinical Applications: Understanding SNPs is crucial in personalized medicine, as specific SNPs can impact an individual's risk of developing related diseases or their response to specific treatments.

Gene networks, also called gene regulatory networks, represent a comprehensive assembly of genes and their interactions dictating the expression levels of genes within a cell or organism [19]. These networks show the connections among genes,

transcription factors, and regulatory molecules and how their interactions affect their expression. They provide insights into complex regulatory mechanisms that control cellular functions such as development and response to environmental triggers. Through mapping these relationships, researchers can better understand how genes work together to sustain biological functions and how variations to the network result in disease. With their comprehensive perspective on gene regulation, gene networks are crucial for understanding molecular pathways in biology and diseases.

For a better understanding of the gene network-based approaches, some datasets will be mentioned: Protein-Protein interaction(PPI) databases such as BioGRID, IntAct, and STRING are valuable for understanding protein functions and cellular processes [23]. Researchers investigate these processes by documenting physical interactions among proteins. Gene expression datasets are derived from microarray or RNA sequencing studies and determine co-expression relationships among genes, which suggest potential regulatory connections [24]. Transcription factor binding datasets obtained from chromatin immunoprecipitation (ChIP) experiments disclose regulatory interactions between transcription factors and their target genes [25]. MicroRNA target databases contain predicted and experimentally validated interactions between microRNAs and target genes, providing insight into post-transcriptional gene regulation [26]. Disease-specific datasets incorporate information on genetic mutations, gene expression alterations, and phenotypic changes. These facilitate investigating how gene networks are disturbed in disease conditions.

Scientists use datasets from diverse experimental strategies to construct and analyze gene networks. Integration and analysis of these diverse datasets are pivotal in constructing comprehensive gene regulatory networks that facilitate unraveling the intricate complexities of gene regulation and cellular processes. Various computational and statistical methods are employed in this endeavor. Network inference algorithms, such as Bayesian networks and mutual information-based methods, leverage gene expression data to infer regulatory interactions between considered genes [27]. Clustering algorithms aid in identifying functional modules within the network by grouping genes with similar expression patterns. Network topology analysis entails studying the structural properties of the network, such as centrality measures and

degree distribution to comprehend its organization and dynamics. Pathway analysis identifies biological pathways enriched with genes of interest, providing crucial insights into the functional roles of gene networks [28]. By combining these methodological approaches, researchers can explain the regulatory relationships within gene networks, uncover vital regulatory mechanisms, and understand how genes coordinate their activities to govern cellular processes. Gene networks thus serve as essential tools in interpreting the molecular mechanisms supporting biological processes and diseases, offering a holistic perspective on gene regulation within biological systems.

Considering all this information, that is a known fact that many common disorders are known to be polygenic today, yet the specific genetic variant combinations remain unidentified [29]. An individual's genetic risk for a particular trait or disease is measured by polygenic risk scores (PRS), commonly calculated by summing the weighted effect sizes of risk alleles at multiple genetic variants [30]. The search for those variants and their combined effects presents significant challenges due to the many possible combinations. Also, incomplete genetic variant information and imprecise effect sizes can affect the accuracy of PRS. Moreover, tagging SNPs instead of causal variants limits the precision of PRS. Given that the current catalog of genetic variants associated with various disorders is incomplete, substantial improvements are necessary to enhance the accuracy of PRS calculations. New methodologies are being developed to address these challenges, such as Bayesian LDpred [31], which incorporates Linkage Disequilibrium (LD), and SBayesR [32], which uses Bayesian multiple regression on summary statistics. Ongoing research aims to improve PRS for disease risk prediction and individualized healthcare strategies by strengthening their clinical utility and dependability.

As mentioned, in gene studies, datasets are generally subjected to statistical analysis. These techniques include population stratification control, association testing, meta-analysis, imputation, mixed-effects models, and data quality control [8] [33] [34] [35] [36] [37]. While these methods have yielded remarkable results, they have limitations: population stratification risks false associations, multiple testing increases false positives, small samples hinder detection, missing heritability suggests

incomplete understanding, focusing on common variants may overlook rare ones, and analyzing gene-environment interactions is complex. The complexity of the analysis and computational cost necessitates exploring new solutions, especially considering the size of the genome data.

Converting genetic sequences into 2D images represents a promising solution to address the challenges posed by multidimensional computations and data complexities in genome studies. These techniques involve representing genetic sequences as visual images, simplifying the computational analysis and interpretation of complex genomic data. Several studies have highlighted the efficacy of this approach, demonstrating that the transformation of genetic sequences into 2D images can lead to acceptable classification accuracies [38]. Researchers have analyzed patient/control genotype data as transformed images using 1D Convolutional Neural Network (1DCNN) and 2D Convolutional Neural Network (2DCNN) architectures. The ability of these neural network algorithms to recognize characteristics and patterns in images allows for accurate genome data classification [19].

High Dimensional Model Representation (HDMR) is a powerful tool for managing and analyzing complex systems. This thesis uses this method to handle the converted 2D images of gene sequences. This method has been extensively applied in engineering, environmental sciences, finance, and biological systems [39]. In gene networks, HDMR presents comprehensive and network-specific gene variant analyses. It is proven that the effectiveness of HDMR as a method for analyzing genetic data and establishes a foundation for future applications of this approach to diploid sequencing data [40].

2. METHODS

2.1 Chaos Game Representation

The Chaos Game Representation approach uses ideas from chaotic systems and nonlinear dynamics to represent genetic sequences visually [20]. An elaborate pattern known as an attractor is produced by the Chaos Game algorithm, which assigns DNA bases to vertices inside a square grid. These attractors reveal the basic structural features inherent in DNA sequences, making it possible to observe the patterns of both localized and global sequences. The CGRs exhibited self-similar patterns recurring at different scales. This similarity highlights gene sequences with repeating motifs. Furthermore, every point on CGR represents a distinct DNA subsequence, making it easier to identify sequence trends.

1. Initialization: In a two-dimensional Cartesian coordinate system, the origin (0, 0) is where the Chaos Game Representation (CGR) begins.

2. Nucleotide Assignment: Adenine (A), thymine (T), guanine (G), and cytosine (C) were assigned specific coordinates. Adenine (A) is linked to (-1, 1) and is positioned in the upper-left quadrant. Thymine (T) is represented by (1, -1) in the lower right quadrant. Guanine (G) occupies (1, 1) and is in the upper-right quadrant. Cytosine (C), found in the lower-left quadrant, was designated as (-1,-1).

3. Iterative Process: The nucleotide sequences of genes were processed iteratively using the CGR algorithm. Like $k-mer$ tables or forward Markov chain analysis, this is frequently performed in reverse order. The actions listed here were performed for each nucleotide. The halfway point is located between the current position and the designated coordinate. The new location is then marked by moving to this halfway point.

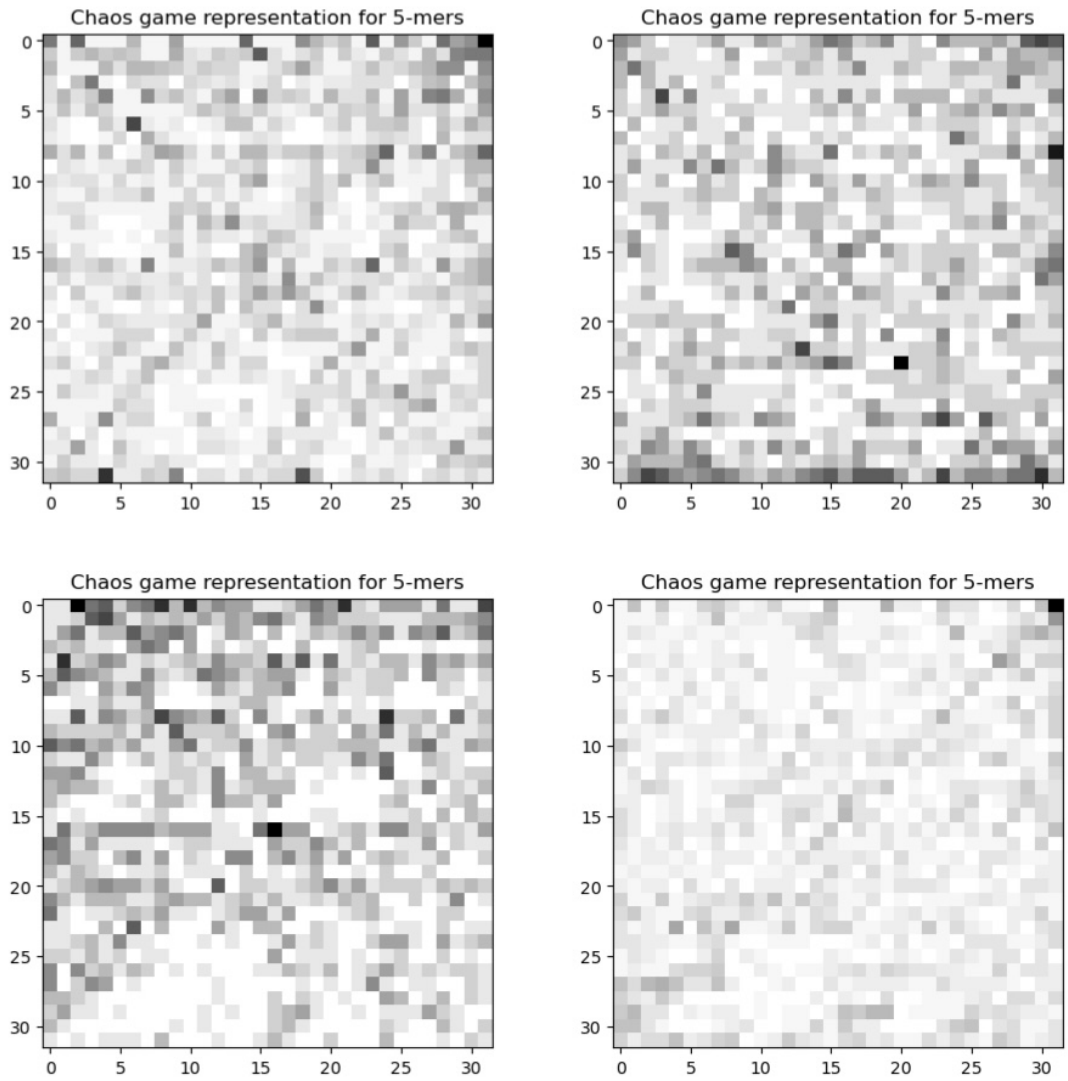

Figure 2.1 : 32×32 CGR images of mTOR gene: GSK3B (top-left), RICTOR (top-right), SLC38A9 (bottom-left), MTOR (bottom-right).

4. Repetitive Process: This procedure is reiterated for each nucleotide in the gene sequence until all nucleotides have been addressed.

An example of some CGR images represented in Figure 2.1 belongs to the mTOR gene network. A subsequence of length in an RNA or DNA sequence is known as a $k-mer$, an essential building block for analyzing and comprehending genomic sequences in computational biology and bioinformatics. DNA sequences of different lengths may be seen owing to the resolution of the $k-mer$ method, which provides information on the composition and structure of the sequences. Furthermore, the correlation between CGR points and sequences shows that mathematical representations of CGR reflect

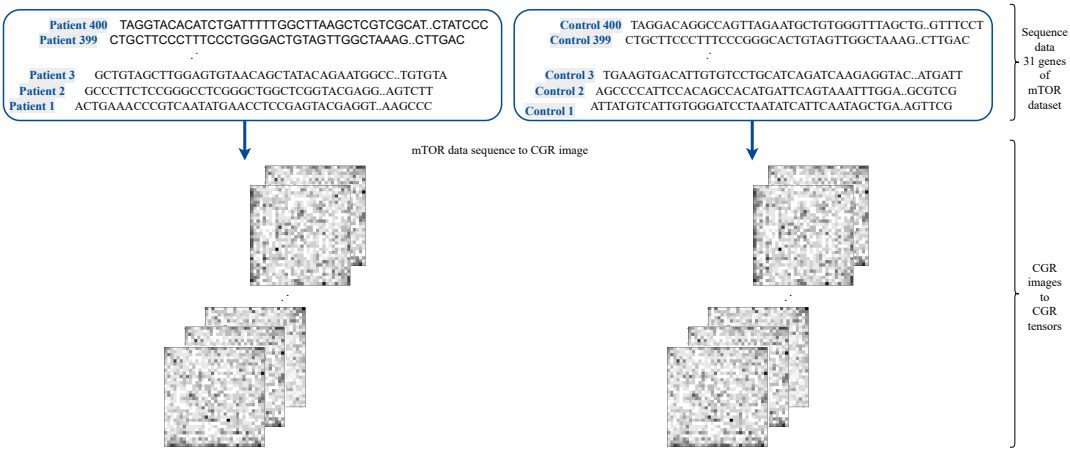

Figure 2.2 : CGR process of genomic sequence data.

features of the underlying DNA sequence, making it easier to analyze sequential and statistical qualities [41].

In our study, we systematically explored a range of $k-mer$ lengths within the range of: $k-mer = \{3,4,5,6,7,8,9\}$. Moreover, evaluated their impact on the analytical outcomes. To achieve this, we assessed the classification accuracy of the control and patient groups. By analyzing the performance of different $k-mer$ lengths, we aimed to identify the optimal parameter that enhances the accuracy of distinguishing between these groups. This comprehensive evaluation involved comparing the classification results to determine how variations in $k-mer$ length influence the overall effectiveness of the analysis.

Following determining the optimal $k-mer$ length, a CGR tensor was constructed through the alignment of the CGR images extracted from all sequences within the patient group. This iterative process was replicated for the control group. Thus, our analysis yielded two distinct CGR tensor classes, representing the patient and control groups.

As represented in Figure 2.2, after the conversion of genomic sequences into FASTA format, CGR images were produced based on the designated $k-mer$ count. The CGR images, corresponding to each respective group, control, and patient, were then aligned in identical sequences to form CGR tensors.

2.2 VARCH

VARCH is a novel method that uses a variant map, which is essential for constructing variant logic and visualizing DNA sequences. This approach counts the frequency of base pairings inside each DNA subsequence, making the visualization process easier to understand. These combinations are then used to create a matrix projected onto a two-dimensional coordinate plane to represent DNA sequences visually. Compared to traditional techniques, VARCH methodology provides a more intuitive and easily interpretable representation of DNA sequence properties [21].

A new logical construct, Variant Logic, was developed to broaden the scope of Boolean functions. It incorporates four fundamental relationships, which signify the primary types of change patterns: 0 to 0, 0 to 1, 1 to 0, and 1 to 1 [42]. The meta-operators represent these relationships ($\{\perp, +, -, \top\}$) in the Variant Logic Constructor (VLC), which corresponds to the mentioned meta change patterns as represented in Table 2.1 [21].

Table 2.1 : The relationships in variant logic construction.

Input	Output	Truth	False	Variant	Invariant	Operator
0	0	0	1	0	1	$\perp$
0	1	1	0	1	0	$+$
1	0	0	1	1	0	$-$
1	1	1	0	0	1	$\top$

As shown in Table 2.2, the four meta-operators may be used in various ways to provide 16 distinct combinations, which provides a more thorough understanding. Each combination provides a strong foundation for examining and extending the logical domain of Boolean functions, which reflects a different application of the meta operators. The VLC may be used to precisely characterize and manipulate logical connections and alter patterns thanks to this extensive collection of combinations.

The primary constraint faced with implementing this method is the inconsistency in the lengths of data sequences. It is a common observation, not just within our investigation but also across various species, that the lengths of DNA sequences display substantial variation when multiple genes are analyzed collectively to yield a comprehensive

Table 2.2 : The relationship between subsets and values.

Name	Value	Base
T_0	$\emptyset$	$\emptyset$
T_1	$N_\perp$	N_C
T_2	N_+	N_A
T_3	N_-	N_T
T_4	$N_\top$	N_G
T_5	$N_\perp + N_+$	$N_C + N_A$
T_6	$N_\perp + N_-$	$N_C + N_T$
T_7	$N_\perp + N_\top$	$N_C + N_G$
T_8	$N_+ + N_-$	$N_A + N_T$
T_9	$N_+ + N_\top$	$N_A + N_G$
T_{10}	$N_- + N_\top$	$N_T + N_G$
T_{11}	$N_\perp + N_+ + N_-$	$N_C + N_A + N_T$
T_{12}	$N_\perp + N_+ + N_\top$	$N_C + N_A + N_G$
T_{13}	$N_\perp + N_- + N_\top$	$N_A + N_C + N_T$
T_{14}	$N_+ + N_- + N_\top$	$N_A + N_T + N_G$
T_{15}	$N_\perp + N_+ + N_- + N_\top$	$N_C + N_A + N_T + N_G$

variant analysis. This inconsistency in data length presents a significant challenge in achieving a reliable and accurate analysis. In the subsequent steps of our study, as we examine the 16 unique combinations and look for accurate matches, we will be partitioning the DNA sequences into equal sub-lengths. This division aims to achieve a more balanced and equal data distribution for the further analysis. Following the methodology adopted in the original study introducing the VARCH method, a single division process will be employed to derive first-degree subsequences of equal length. By doing so, we aim to overcome the inherent challenges of variable data length, thereby improving the reliability and accuracy of our analytical outcomes.

The connection may be written as in Equation 2.1, if the length of a DNA sequence is represented by N, the length of a subsequence by Len, and the number of subsequences by m.

$$N = Len \times m \tag{2.1}$$

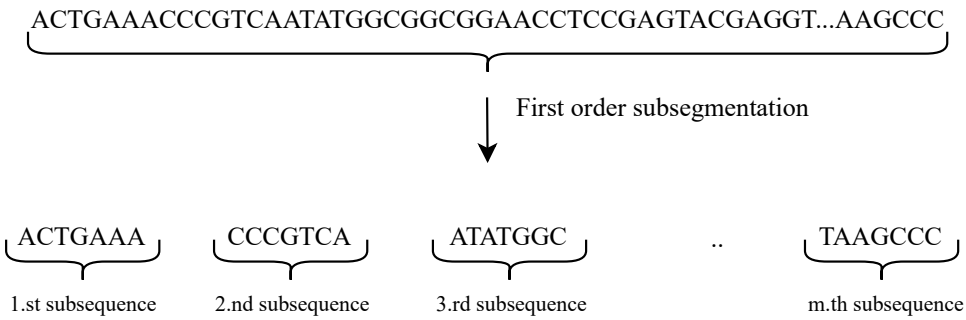

Figure 2.3 : Subsegmentation of the genomic sequence data.

Every participant in our study has had their gene sequences subjected to this procedure, which guarantees that the total length of the subsequence is divided into the subsegment's length, as seen in Figure 2.3. This methodological technique makes accurate comparisons and thorough analyses possible by enabling a uniform and consistent study across all DNA sequences.

After the initial step of obtaining the subsequences from the main dataset, we can proceed to the calculation of the matrix, Mat(A). This is achieved by utilizing the corresponding subset numbers associated with each subsequence. It is important to note that there are 16 unique configurations, denoted from T_0 to T_{15}. Additionally, we have a number of subsequences derived from the original data.

Considering these variables, the resulting matrix, $\text{Mat}(A) \in \mathbb{R}^{m\times 16}$ will have dimensions of $m \times 16$ a relationship clearly illustrated in Equation 2.1. This configuration was chosen because it allows for a comprehensive and complete data representation. It successfully captures the full range of possible combinations within the subsequences, thus providing a solid foundation for further analysis and interpretation of the data.

$$\text{Mat}(A) = \begin{bmatrix} T_0^1 & T_1^1 & \cdots & T_j^1 \cdots & T_{15}^1 \\ T_0^2 & T_1^2 & \cdots & T_j^2 \cdots & T_{15}^2 \\ \vdots & \vdots & \ddots & \vdots & \\ T_0^i & T_1^i & \cdots & T_j^i \cdots & T_{15}^i \\ \vdots & \vdots & \ddots & \vdots & \\ T_0^m & T_1^m & \cdots & T_j^m \cdots & T_{15}^m \end{bmatrix} \tag{2.2}$$

Equation 2.2 represents how matrix A, comprehensively compiles all potential combinations. The task at hand is to choose a pair of T_α and T_β that will be used to construct the visualization matrix Mat(B) based on the accuracy findings. The challenge lies in determining which alpha (α) and beta (β) combination most effectively encapsulates the patterns and information embedded within the DNA sequence.

Given the complexity of this task, a comprehensive approach was taken. Every possible pair, every Mat(B) combination, has been created. The performance of these matches was then evaluated for classification accuracy, providing a quantifiable measure to determine their effectiveness.

After careful analysis that examined accuracy, F1 score, recall, and loss results, the T_α and T_β pairing that provided the highest performance was selected. Once this pair was identified, the process moved on to the next step. This involved recording the frequency information into Mat(B) as Equation 2.3 represented. The information on how many pairs (T_α ,T_β) are present in the corresponding elements can be called frequency information, the order, and the number of matches associated with this selected (T_α ,T_β) pair in the resulting Mat(A) are also documented. This comprehensive approach allows us to capture the most accurate and detailed representation of the gene sequence.

$$\mathrm{Mat}(B) = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1h}\cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2h}\cdots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots & \\ b_{i1} & b_{i2} & \cdots & b_{ih}\cdots & b_{im} \\ \vdots & \vdots & \ddots & \vdots & \\ b_{m1} & b_{m2} & \cdots & b_{mh}\cdots & b_{mm} \end{bmatrix} \tag{2.3}$$

where (T_α^i ,T_β^h) coordinate's frequency value is b_{ij}. The coordinate (T_α^i ,T_β^h) denotes that the two configurations align with specific elements within the matrix Mat(A) as (T_α^1 ,T_β^1), (T_α^m ,T_β^m) , ... , (T_α^i ,T_β^h) while $1 \leq i \leq m$, $1 \leq h \leq m$ [21].

For an example, as shown in Figure 2.4 the mTOR gene sequence VARCH images, represented with heatmap data visualizationn tool, varying with the T_α and T_β values show that there is a symmetry between the four bases and 16 configurations.

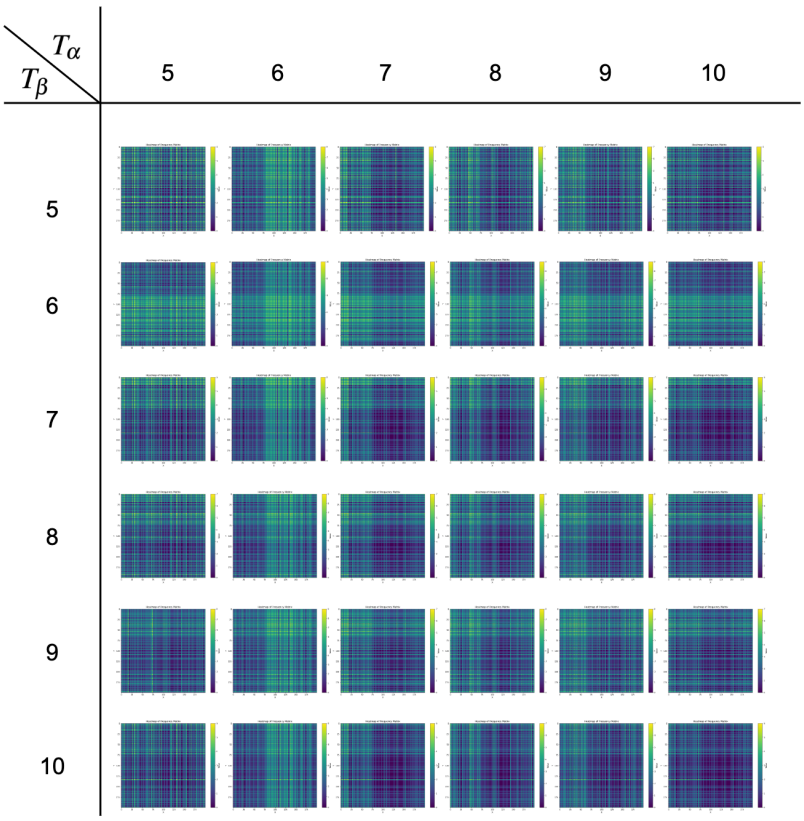

Figure 2.4 : VARCH images of varying T_α and T_β values ($Len = 10$, $m = 200$).

Our primary focus is to maximize the amount of vital information that can be extracted from this step. Doing so can significantly support generalization performance, leading to more accurate and reliable results that can further our understanding. In light of this, we have decided to limit the gene sequence length. This decision ensured that the sequence length matched the shortest gene sequence's length in our mTOR dataset.

The equation that has been the subject of our study is represented as Equation 2.1 and involves two variable parameters, which must be adjusted during the calculation process. These parameters are specifically referred to as parameters *Len* and *m*. The pivotal element within this equation is the *Len* parameter, which is the subsequence length parameter. Its significance and integral role in determining the outcomes of our calculations were thoroughly proven and elaborated upon in our work. This finding is demonstrated through the accuracy of a Support Vector Machines classification, discussed in Section 4. This section comprehensively analyzes the *Len* parameter's role and impact on the outcome, further underlining its importance in our research.

After performing a series of computations, we generated two tensors of size $[m \times m]$ for the patient and control groups. These were created by collating the VARCH images. It is important to note that the size of these tensors varies according to the content of the datasets used in each study. This variation is solely determined by the number of subsequences examined in the study. The VARCH tensor is obtained by aligning the frequency matrices B, Mat(B), which we obtained, with the data from the patient and control groups. These tensors will be input into the HDMR method similarly to the CGR tensor.

2.3 High Dimensional Model Representation

High Dimensional model representation (HDMR) is a technique used to predict complex relationships between input and output variables in high dimensional systems [22] [39]. High Dimensional Model Representation allows for the representation of a multivariate function $f(x_1, x_2, \ldots, x_N)$ as a finite sum of functions that have fewer independent variables and are mutually orthogonal. This method, which assumes minimal high-order interaction effects, breaks down the output function into hierarchical components. This allows it to represent the independent or combined effects of the input variables. The computational efficiency of HDMR remains constant regardless of the size of the function by utilizing sampling methods, such as Monte Carlo integration, to handle dispersed data points effectively. Error analysis and comparison with the test datasets are used for validation and assessment. The procedure comprises choosing basis functions, determining sample sizes, and using predetermined computational techniques. Overall, a high dimensional multivariate representation of functions offers a methodical and efficient option.

Finding patterns and demonstrating relationships in genetic operational analysis research is computationally expensive owing to its large dimensionality and complicated challenges. However, the HDMR method should be considered a solution for these problems in gene research [43] [44]. The HDMR operates by splitting a multivariate function into a sum of hierarchical functions [22].

$$f(x_1,\ldots,x_N) = f_0 + \sum_{i_1=1}^{N} f_{i_1}(x_{i_1}) + \sum_{\substack{i_1,i_2=1 \\ i_1<i_2}}^{N} f_{i_1 i_2}(x_{i_1},x_{i_2}) + \cdots + f_{12\ldots N}(x_1,\ldots,x_N) \quad (2.4)$$

This expansion represents a finite summation involving constant, univariate, bivariate, and up to N-variate functions. On the right-hand side of the expansion, the expression includes a singular constant function (f_0), N univariate functions $(f_1, f_2, \ldots, f_N)$, $\frac{N(N-1)}{2}$, bivariate functions $(f_{12}, f_{13}, \ldots, f_{1N}, f_{23}, f_{24}, \ldots, f_{2N}, \ldots, f_{N,N_1})$, and continues in this manner up to a total of 2^N functions. These terms are categorized as the constant term, univariate terms, bivariate terms, and extending to N-variate terms. Each univariate term quantifies the influence of its corresponding independent variable within the function's structure, whereas each bivariate term captures the interactive effects of the involved independent variables. By selecting a suitable subset of terms from this expansion, one can approximate the multivariate function $f(x_1,\ldots,x_N)$ with the desired degree of precision.

For example, considering solely the constant term and ignoring all higher-order terms yields a zeroth-order approximation. This approximation corresponds to the average value of the function over the hyperprism $[a,b]^N$. The hyperprism $[a,b]^N = [a_1,b_1] \times \cdots \times [a_N,b_N]$ is defined as the Cartesian product of the individual intervals.

When the function being approximated is constant, the zeroth-order approximation accurately represents this value. In cases where terms involving two or more variables are disregarded, the result is a first-order approximation. Similarly, excluding terms that involve three or more variables in the Equation 2.4 produces a second-order approximation. This approach enables the multivariate function to be approximated by focusing on functions that involve fewer variables.

If HDMR and its components will be theoretically examined the weight function is defined as a completely multiplicative function:

$$W(x_1,\ldots,x_N) = \prod_{i=1}^{N} w_i(x_i) \quad (2.5)$$

Each factor $w_i(x_i)$is chosen so that its integral over the interval $[a_i,b_i]$ is equal to 1:

$$\int_{a_i}^{b_i} w_i(x_i)\, dx_i = 1 \tag{2.6}$$

If this integral equals a constant c, then the corresponding factor is scaled by $1/c$ to ensure that Equation 2.6 is satisfied.

Each HDMR component is such that $x_i \in \{x_{i_1}, x_{i_2}, \ldots, x_{i_k}\}$,

$$\int_{a_i}^{b_i} w_i(x_i) f_{i_1 i_2 \ldots i_k}(x_{i_1}, x_{i_2}, \ldots, x_{i_k})\, dx_i = 0; \quad i \in \{i_1, i_2, \ldots, i_k\} \tag{2.7}$$

For each HDMR component to be uniquely defined, it must adhere to the "vanishing under integration" condition, which requires the Equation 2.7 is sufficient for the unique determination of HDMR components.

Given these conditions, the HDMR components can be uniquely identified. To present these components more systematically, we first need to introduce some definitions. Let dx^i represent the differential operator with respect to all variables except x_i, and dx^{ij} denote the differential operator with respect to all variables other than x_i and x_j. These operators are defined as follow in Equation 2.8, 2.9 and 2.10:

$$dx = dx_1 \cdots dx_N \tag{2.8}$$

$$dx^i = dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_N \tag{2.9}$$

$$dx^{ij} = dx_1 \cdots dx_{i-1}\, dx_{i+1} \cdots dx_{j-1}\, dx_{j+1} \cdots dx_N \tag{2.10}$$

Employing these definitions, the components of the multivariate function $f(x_1, \ldots, x_N)$ can be represented as in Equation 2.11 :

Constant Component f_0:

$$f_0 = \int_{a_1}^{b_1} \cdots \int_{a_N}^{b_N} \prod_{k=1}^{N} w_k(x_k) f(x_1, \ldots, x_N)\, dx \tag{2.11}$$

Univariate Components $f_i(x_i)$:

$$f_i(x_i) = \int_{a_1}^{b_1} \cdots \int_{a_{i-1}}^{b_{i-1}} \int_{a_{i+1}}^{b_{i+1}} \cdots \int_{a_N}^{b_N} \prod_{\substack{k=1 \\ k \neq i}}^{N} w_k(x_k)\, f(x_1, \ldots, x_N)\, dx^i - f_0 \tag{2.12}$$

Bivariate Components $f_{ij}(x_i, x_j)$:

$$\begin{aligned} f_{ij}(x_i, x_j) = \int_{a_1}^{b_1} \cdots \int_{a_{i-1}}^{b_{i-1}} \int_{a_{i+1}}^{b_{i+1}} \cdots \int_{a_{j-1}}^{b_{j-1}} \int_{a_{j+1}}^{b_{j+1}} \cdots \int_{a_N}^{b_N} \prod_{\substack{k=1 \\ k \neq i,j}}^{N} w_k(x_k) \times \\ \times f(x_1, \ldots, x_N)\, dx^{ij} - f_0 - f_i(x_i) - f_j(x_j) \end{aligned} \tag{2.13}$$

Each component is computed accordingly. The zeroth, first, and generally k-th order HDMR approximations are expressed as in the Equation 2.14, 2.15 and 2.16:

$$s_0 = f_0 \tag{2.14}$$

$$s_1 = s_0 + \sum_{i=1}^{N} f_i(x_i) \tag{2.15}$$

$$\vdots$$

$$s_k = s_{k-1} + \sum_{\substack{i_1, i_2, \ldots, i_k = 1 \\ i_1 < i_2 < \cdots < i_k}}^{N} f_{i_1 i_2 \cdots i_k}(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) \tag{2.16}$$

Equation 2.7 shows that the HDMR components can be shown to be orthogonal to each other. According to this equality, for $1 \leq i_1 < i_2 < \cdots < i_k \leq N, \quad 1 \leq j_1 < j_2 < \cdots < j_l \leq N, \quad 1 \leq k < l \leq N$,

$$\int_{a_1}^{b_1} \cdots \int_{a_N}^{b_N} \prod_{k=1}^{N} w_k(x_k)\, f_{i_1 i_2 \cdots i_k}(x_{i_1}, x_{i_2}, \ldots, x_{i_k})\, f_{j_1 j_2 \cdots j_l}(x_{j_1}, x_{j_2}, \ldots, x_{j_l})\, dx = 0 \tag{2.17}$$

the Equation 2.17 must hold true, indicating that all HDMR components are mutually orthogonal. This orthogonality simplifies the calculation of the components and allows for the definition of an inner product. The inner product can be expressed as follows:

$$(f_{i_1 i_2 \cdots i_k}, f_{j_1 j_2 \cdots j_l}) = \int_{a_1}^{b_1} \cdots \int_{a_N}^{b_N} \prod_{k=1}^{N} w_k(x_k)\, f_{i_1 i_2 \cdots i_k}(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) \times \\ \times f_{j_1 j_2 \cdots j_l}(x_{j_1}, x_{j_2}, \ldots, x_{j_l})\, dx \tag{2.18}$$

Using this inner product, a norm can be defined for a function $f(x_1, \ldots, x_N)$ that is integrable in the square as follows:

$$\|f_{i_1 i_2 \cdots i_k}\|^2 = (f_{i_1 i_2 \cdots i_k}, f_{i_1 i_2 \cdots i_k}) \tag{2.19}$$

If the Equation 2.4 is multiplied on both sides by itself and by the weights, and then integrated over the corresponding ranges for all variables, the Equation 2.20 is obtained with the zeros resulting from the orthogonality condition in Equation 2.17.

$$\|f\|^2 = \|f_0\|^2 + \sum_{i=1}^{N} \|f_i\|^2 + \sum_{\substack{i,j=1 \\ i<j}}^{N} \|f_{ij}\|^2 + \cdots + \|f_{12\cdots N}\|^2 \tag{2.20}$$

Thus, it is concluded that the square of the norm of $f(x_1, \ldots, x_N)$ is equal to the sum of the squares of the norms of its components. If both sides of this equality are divided by $\|f\|^2$, we get the Equation 2.21:

$$1 = \frac{\|f_0\|^2}{\|f\|^2} + \frac{1}{\|f\|^2} \sum_{i=1}^{N} \|f_i\|^2 + \frac{1}{\|f\|^2} \sum_{\substack{i,j=1 \\ i<j}}^{N} \|f_{ij}\|^2 + \cdots + \frac{1}{\|f\|^2} \|f_{12\cdots N}\|^2 \tag{2.21}$$

By appropriately making cuts at the desired levels, we get these equations:

$$\begin{aligned} \sigma_0 &= \frac{\|f_0\|^2}{\|f\|^2}, \\ \sigma_1 &= \frac{1}{\|f\|^2} \sum_{i=1}^{N} \|f_i\|^2 + \sigma_0 \\ &\vdots \\ \sigma_N &= \frac{1}{\|f\|^2} \|f_{12\cdots N}\|^2 + \sigma_{N-1} \end{aligned} \tag{2.22}$$

Here, σ_i is referred to as the i-th order summability measure. These measures form a well-ordered sequence that takes values between 0 and 1. As can be seen in the

Equation 2.23, the closer the value of σ_i is to 1, the closer the approximation is to the actual result.

$$0 \leq \sigma_1 \leq \sigma_2 \leq \cdots \leq \sigma_N = 1 \tag{2.23}$$

To summarize the preceding information, Equation 2.4 demonstrates that the HDMR method can be effectively applied to N-dimensional data. In the context of this thesis, the N-dimensional tensors obtained through the visualization steps of the VARCH and CGR methods were subjected to the HDMR method for further analysis.

However, it should be noted that the objective of this thesis was not merely to reconstruct the tensors as mentioned above in the same N-dimensional format. Instead, the primary goal was to utilize the inherent functionalities of the HDMR method for data decorrelation, dimension reduction and also feature extraction. These functionalities are of significant importance for data analysis and representation and can significantly contribute to the clarity and comprehensibility of the findings. Therefore, we employed only zeroth and first order components of HDMR was implemented in this thesis, providing a comprehensive overview of the process and its applications. This implementation not only facilitates the understanding of the data but also enhances the method's potential in the context of data representation and interpretation.

After calculating the zeroth-order and first-order terms, we will only vertically concatenate the obtained first-order components f_1, f_2, and f_3 to create our feature matrix for patient and control groups. We utilized the zeroth-order term only during the computation of the first-order components. Even with identical mTOR datasets, two approaches, CGR and VARCH, produce tensors with different dimensions. Calculations will be made using the following equations. First, the zeroth-order component is determined as in Equation 2.26,

$$f(x) \approx f_0 + \sum_{i=1}^{n} f_i(x_i) \tag{2.24}$$

$$f_0 = \frac{1}{n_1 n_2 n_3} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} H_{ijk} \tag{2.25}$$

$$f_1 = \frac{1}{n_2 n_3} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} H_{ijk} - f_0 \tag{2.26}$$

$$f_2 = \frac{1}{n_1 n_3} \sum_{i=1}^{n_1} \sum_{k=1}^{n_3} H_{ijk} - f_0 \tag{2.27}$$

$$f_3 = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} H_{ijk} - f_0 \tag{2.28}$$

where n_1, n_2 and n_3 are the dimension of the CGR tensor under consideration. The first-order components calculations are represented in Equations 2.26, 2.28, and 2.28 [22],

First-order hierarchical functions f_1, f_2 and f_3 were calculated by subtracting the contribution of lower-order terms from that of higher-order terms. From now on we are going to call f_1, f_2 and f_3 functions as $\vec{h_1}, \vec{h_2}$ and $\vec{h_3}$ vectors for clarification. Finally, the features obtained from the hierarchical functions are concatenated into a single feature vector for further analysis. Vectors $\vec{h_1}, \vec{h_2}$ and $\vec{h_3}$ were obtained for each individual, corresponding to their gene sequences. Depending on the visualization method used, the information contained within these tensors, that is, their dimensions, varies. However, the vectors obtained from the two methods were concatenated vertically and recorded in a single column for each individual's information, as illustrated in Equation 2.29,

$$\text{Feature Matrix} = \begin{bmatrix} \vec{h}_{1,1} & \vec{h}_{1,2} & \cdots & \vec{h}_{1,400} \\ \vec{h}_{2,1} & \vec{h}_{2,2} & \cdots & \vec{h}_{2,400} \\ \vec{h}_{3,1} & \vec{h}_{3,2} & \cdots & \vec{h}_{3,400} \end{bmatrix} \tag{2.29}$$

where $\vec{h}_{i,j}$ contains $i = \{1,2,3\}$ and $j = \{1,2,..,400\}$, while there are 400 individuals for each patient and control groups. A feature matrix, a design or predictor matrix, is a structured data representation used in statistical modeling and machine learning. Each row typically represents a sample, and each column represents a feature. The value of each characteristic for a specific sample is in a cell within the matrix. This arrangement

is vital for computational analysis, modeling tasks, training machine learning models, and conducting statistical analyses. HDMR method outcomes two feature matrices: one for the patient group and one for the control group.

2.4 Support Vector Machines

Support Vector Machines is a machine learning technique widely utilized in classification problems constructed over high-dimensional feature spaces. This method falls under supervised learning algorithms primarily used for regression and classification tasks, which are fundamental to data science and machine learning.

The SVM algorithm determines the best hyperplane that divides various classes in the feature space. The aim is to maximize the margin, the distance between the hyperplane, and the nearest data points in each class. By increasing this margin, SVM algorithm improve generalization performance when categorizing fresh data points. SVM algorithm is well known for its capacity to handle high dimensional data and solve challenging classification tasks [45].

The primary purpose of this study is to classify the two groups by using information about the patient and control groups. In the dataset, data from both the patient and control groups were labeled and utilized for training purposes. Subsequently, the model's performance in classifying new data was evaluated by examining information from various individuals within the same dataset. The upmost importance in this process is the proper selection of an appropriate kernel function [46].

This thesis focused on a classification task involving two distinct and clearly defined classes. These classes are represented by the binary outcomes 1 and 0, as the ideal classification with 100% accuracy illustrated in the Figure 2.5 which represent control and patient classifications as, respectively, in the context of this thesis.

The choice of the kernel for this classification task was crucial, and after careful deliberation, the linear kernel was selected. This decision was primarily based on the inherent assumption of linear separability in the dataset of the linear kernel. This assumption aligns with the nature of the data used in this study. Furthermore, the

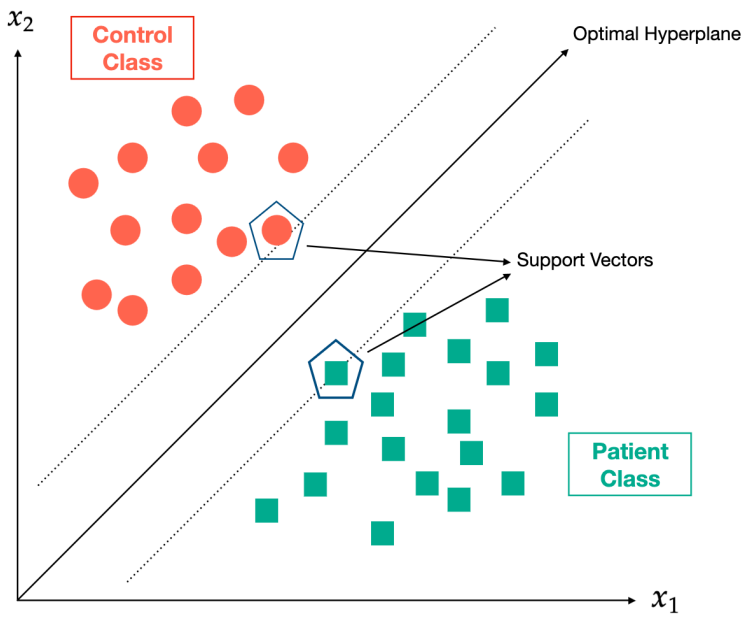

Figure 2.5 : Binary classification using SVM linear kernel.

optimal kernel was selected to enhance our model's generalization capabilities for unseen data, ensure a high accuracy level, and maintain computational efficiency.

The input data for this study can be represented using two feature vectors: $\vec{x}$ and $\vec{y}$. The calculation of linear kernel K exists between these two vectors, as in Equation 2.30,

$$K(\vec{x},\vec{y}) = \vec{x}^T \vec{y} \tag{2.30}$$

This dot product operation is utilized to calculate the similarity between feature vectors. Consequently, larger dot product values indicate higher degrees of similarity between feature vectors. Evaluating the trained model's performance involves assessing its generalization capabilities and accuracy levels. This evaluation was conducted using the training and test datasets subjected to various ratio conditions. The classification of these datasets into imbalanced and balanced categories was performed based on the detailed explanation provided in Section 4.

3. IMPLEMENTATION

3.1 mTOR Dataset

This study used mTOR [47] pathway genes as a dataset, which was selected based on the KEGG database (https://www.genome.jp/kegg/). The genomic sequences of these pathway genes were retrieved from the GRCh37 human genome database, using their genomic coordinates recorded in the NCBI database (https://www.ncbi.nlm.nih.gov/projects/genome/guide/human/index.shtml). Specifically, the dataset consisted of 400 instances each from the patient and control groups, strategically selected to facilitate investigations across different research disciplines. This dataset is a foundational resource for endeavors to refine analytical methodologies, uncover underlying patterns, and foster advancements in cancer research, metabolic disorders, neurodegenerative diseases, immunology, and developmental biology [40].

mTOR, a mammalian Target of Rapamycin, is a critical protein kinase that controls various cellular processes, such as cell growth, proliferation, metabolism, and survival [48]. It acts as a central regulator of signaling pathways in response to environmental cues, such as nutrients, growth factors, and stressors [49]. mTOR also coordinates cellular responses to nutrient availability and energy status changes. Dysregulation of mTOR signaling has been associated with numerous human diseases, including cancer, metabolic disorders, neurodegenerative diseases, and aging. Because of its crucial role in these pathways and diseases, mTOR is an important target for therapeutic interventions in various medical fields.

Studies involving mTOR span diverse fields of biomedical research owing to its crucial role in controlling cellular processes. The role of mTOR in tumor development and progression has been extensively studied, focusing on mTOR inhibitors as potential anticancer agents [50]. mTOR dysregulation is also associated with metabolic disorders, such as obesity and type 2 diabetes, prompting research into mTOR-targeted

therapies for these conditions [51]. The role of mTOR in neuronal function and survival is under investigation in neurodegenerative diseases such as Alzheimer's and Parkinson's, offering insights into potential therapeutic strategies. Furthermore, the association between mTOR and the aging process has sparked studies on anti-aging interventions. In immunology, mTOR signaling affects immune cell function and response, with implications for immune-related diseases and immunotherapy. Finally, developmental biology studies have investigated mTOR's role of mTOR in embryogenesis, tissue regeneration, and developmental disorders. Thus, mTOR research covers diverse areas and continues to reveal its multifaceted functions and implications for health and disease.

mTORC1 and mTORC2 are distinct protein complexes within the mTOR signaling pathway. mTORC1 primarily promotes cell growth and proliferation by stimulating protein synthesis and inhibiting autophagy. It responds to nutrients and growth signals. mTORC2 regulates cell survival and metabolism. Both complexes are critical for cellular homeostasis, and their dysregulation is linked to diseases such as cancer and metabolic disorders. Understanding their roles is vital for developing targeted therapies. Maintaining optimal activity of mTORC1 and mTORC2 is crucial for sustaining metabolic homeostasis and preventing diseases [52].

3.2 Numerical Experiments

3.2.1 Balanced dataset

A balanced dataset is one in which each class or category has approximately the same number of samples. The distribution of classes was approximately equal, without a significant skew towards any class. For instance, in a binary classification problem, a balanced dataset has an approximately equal instance for each class. In multiclass classification, each class has a similar number of samples. Balanced datasets are preferred for machine learning tasks because they prevent biases towards the majority class, ensuring that the model learns from ample examples of each class, resulting in more accurate and reliable predictions.

Balanced datasets are essential in machine learning to minimize biases, improve model performance, promote fairness, aid in effective learning, and enhance interpretability. By providing an equal representation of each class, these datasets prevent models from favoring the majority class, leading to more accurate predictions. This balance enables fair treatment of all classes, effectively allowing models to learn underlying patterns. Ultimately, balanced datasets contribute to the reliability and effectiveness of machine learning models in various domains.

In this thesis, conforming to the balanced dataset practice, 400/400 patients were deployed as control training and test data. For imbalanced and balanced datasets, we used half of the data for testing and the other half for training.

3.2.2 Imbalanced dataset

An imbalanced dataset is one in which the class distribution is skewed, meaning that one class has significantly more instances than the others. This imbalance can challenge machine learning algorithms, especially in classification tasks, as they may struggle to predict minority classes or classes accurately. Consequently, models trained on such datasets may display biases towards the majority class, leading to suboptimal performance and decreased predictive accuracy.

Imbalanced datasets are particularly relevant in gene analysis studies. These datasets often show imbalanced distributions, in which specific genetic variations may be much rarer than others. This imbalance could result in biased models that overlook important genetic markers if not adequately addressed. Accurate classification of genetic variations is critical for understanding their role in disease susceptibility and treatment response. The reliable classification of genetic variants is essential in personalized medicine and genomic healthcare, where treatment decisions are guided by genetic information. Addressing this issue can involve resampling methods, class weighting, or algorithmic adjustments to mitigate the impact of class imbalance and improve the model performance.

In this context, we aimed to use a dataset of 400 patients and 400 control individual data to minimize the computational cost and maximize the classification accuracy. This

will facilitate advances in precision medicine and genomic research. Data selection for this section consists of 500 instances distributed as 400 controls and 100 patients.

4. RESULTS

In this section, we will examine the results of our analysis using the mTOR dataset, whose specifications were provided in Section 3. We will focus on the classification accuracy performances obtained through the CGR, VARCH, and HDMR methods, facilitated by an SVM algorithm. The entire process is shown in Figure 4.1.

As outlined in previous sections, our objective was to minimize computational costs while maximizing accuracy across the methods. To this end, we systematically explored the variations in different parameters for each method and elucidated the rationale behind our parameter selections alongside the results. Furthermore, we discussed the impact of using imbalanced and balanced datasets on these results. By providing a comprehensive overview of the parameter adjustments and dataset considerations, we aimed to shed light on the optimization process and the factors influencing classification accuracy. This analysis serves to enhance our understanding of the performance of each method and their applicability in variant analysis, ultimately contributing to advancements in the field.

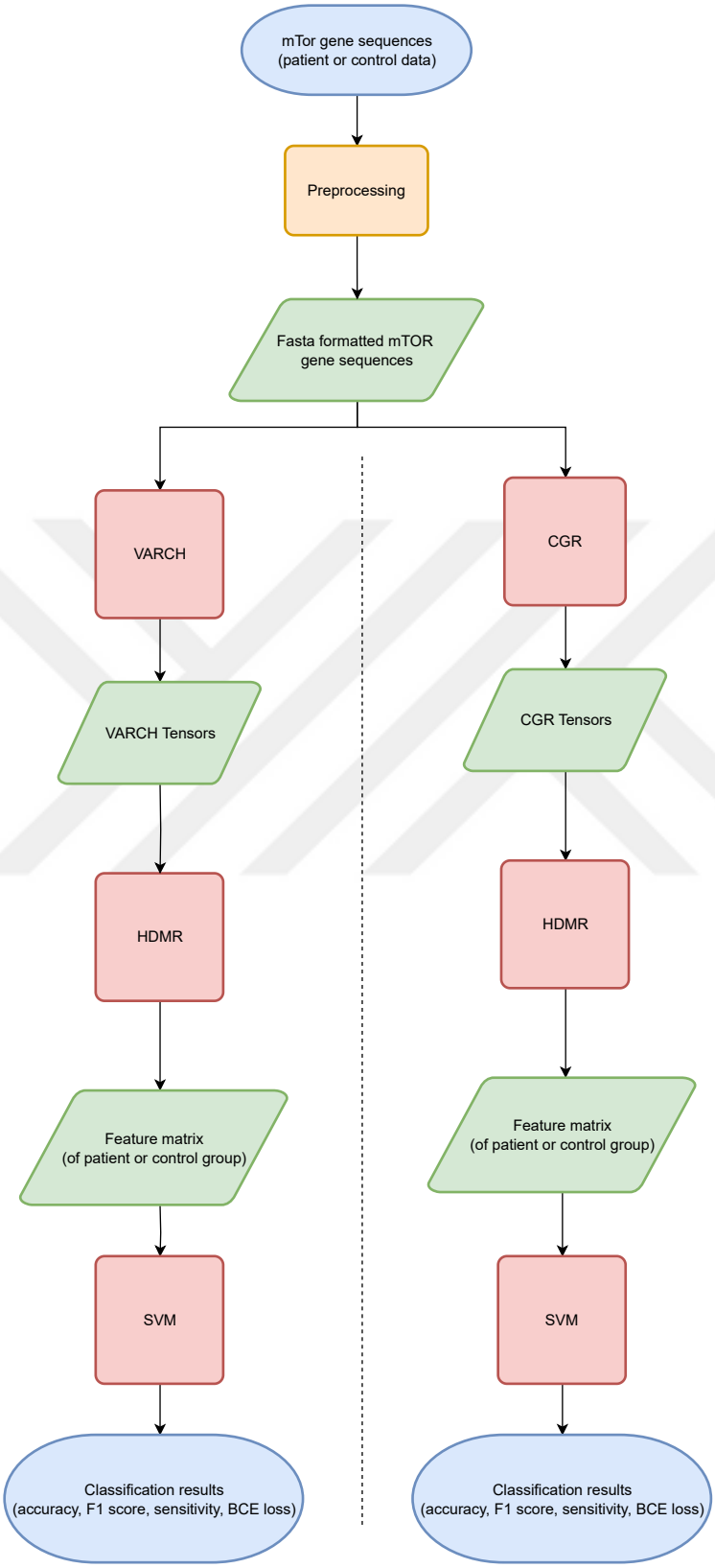

Figure 4.1 : Flowchart of the proposed Gene Network Analysis Process.

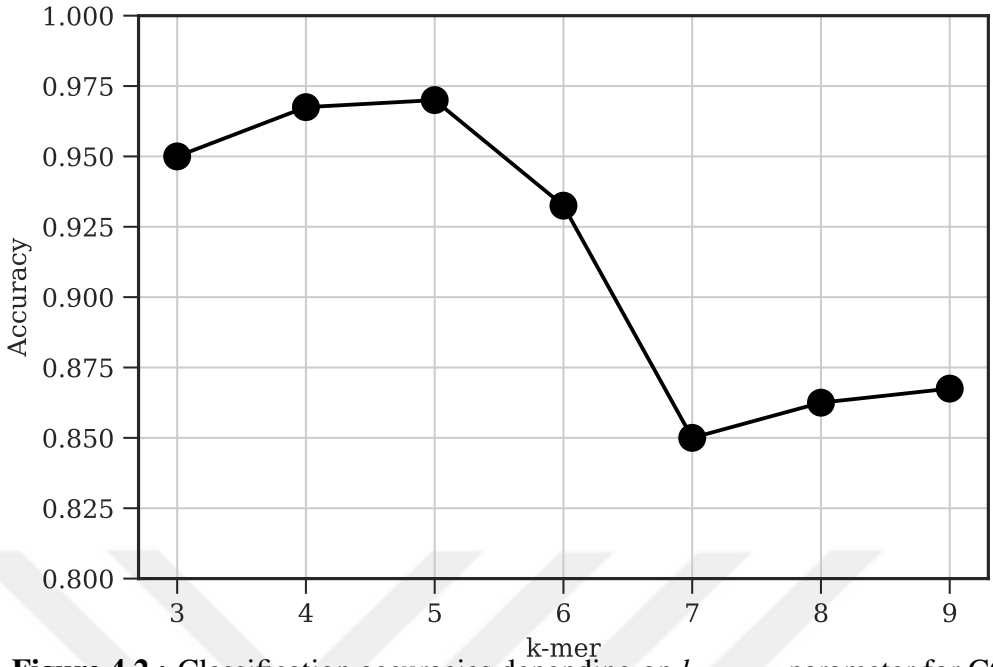

Figure 4.2 : Classification accuracies depending on $k-mer$ parameter for CGR and HDMR (400 patient/control data).

The $k-mer$ parameter in the CGR method acts as the resolution element. As represented in Figure 4.2, the accuracy changes when applying the CGR and HDMR methods to data with a linear kernel SVM, depending on the $k-mer$ parameter. If we view the $k-mer$ parameter as a resolution, we might expect the data within the images to become more abundant as the resolution increases. However, we observed an increase in accuracy up to $n = 5$, followed by a decline. This could be attributed to two reasons. First, smaller $k-mer$ values result in a shorter, more abundant series of $k-mer$s, providing higher-resolution CGR images in more detail. However, this high resolution can make the interpretation challenging. On the other hand, larger $k-mer$ values yield a longer, less abundant series of $k-mer$s, leading to lower-resolution CGR images that are easier to interpret. Therefore, in our experiments with the mTOR genomic dataset, an ideal $k-mer$ parameter of 5 was observed. All future experiments will be performed using this parameter as a standard.

Similarly, the VARCH method requires several trials to determine optimal parameters. One challenge is the variability in gene sequence length within the mTOR dataset. This is because sequence data are stored as matrices, where each gene sequence's data are

stored in a matrix row. Hence, all the gene sequences must be of the same length. The shortest sequence length among the gene sequences determines the sequence length, meaning the *Len* parameter is dictated by the shortest genome sequence in our mTOR dataset. Although this varies depending on the dataset, it often may lead to a significant data loss. After determining this parameter, we can analyze the subsequence length and the number of subsequences. The subsequence length is similar to the $k-mer$ parameter in the CGR method, which affects resolution. The predefined *Len* parameter and subsequence length determine the number of subsequences.

We evaluated the VARCH method's output as input for the HDMR method, which could increase classification accuracy, considering trade-offs with computational cost. Experiments were limited to the shortest gene sequence length, 1600 in this study, and various parameter settings, as shown in Figure 4.3. For the VARCH method, several trials were conducted to determine the optimal parameters. Therefore, depending on the dataset, the shortest gene sequence dictates the *Len* parameter, which can lead to substantial data loss.

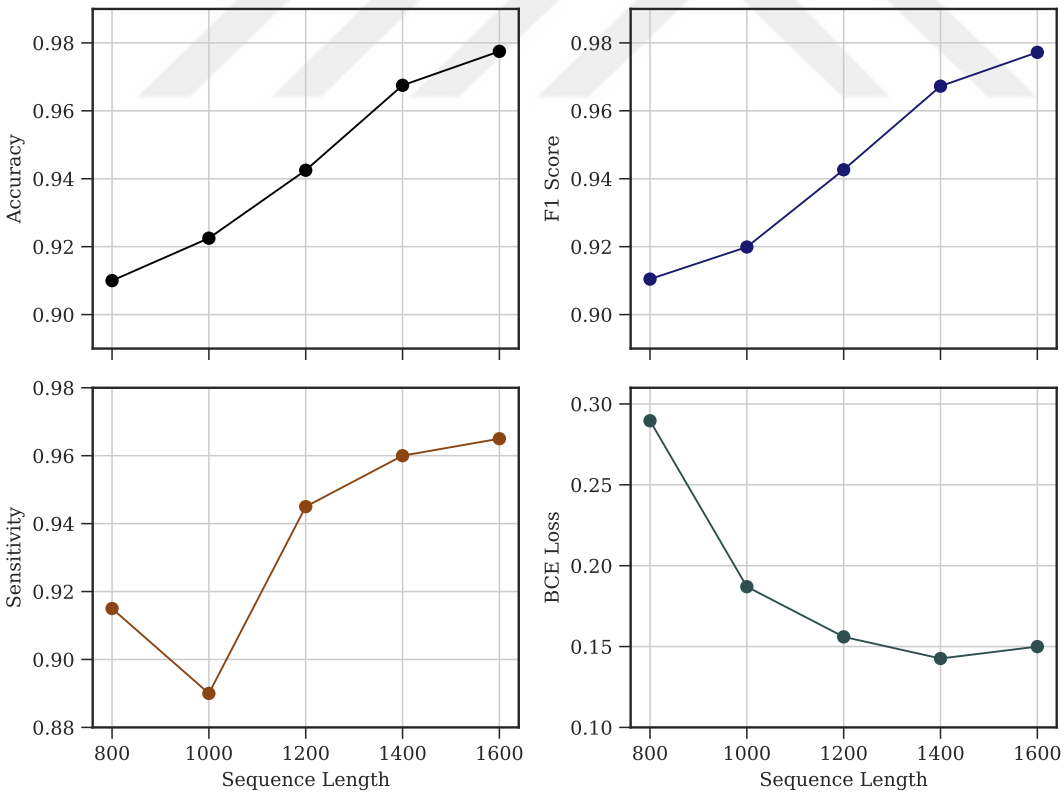

Figure 4.3 : Varying effects of VARCH's sequence length on accuracy, F1 score, recall and BCE loss.

As expected, increasing the total sequence length reduced data loss and enabled training with larger datasets, leading to more accurate results. A detailed analysis reveals that accuracy and F1 score positively correlate with sequence length. Similarly, as anticipated, the BCE loss decreased with increasing sequence length.

Since the sequence length is set, the other parameter, the number of subsequences, is determined by the predefined *Len* parameter and the subsequence length. Experiments were conducted with *Len* values of 40, 20, and 10 to increase the resolution and discuss the results. However, because the method relies on searching for the 16 combinations shown in Table 2.2 within each subsequence, the subsequence lengths cannot be shorter than the combinations. Thus, we selected ten as the minimum subsequence length. The classification accuracy, F1 score, sensitivity (recall), and binary cross-entropy loss results are recorded in Table 4.1,

Table 4.1 : VARCH results for varying *Len* and *m* parameters ($T_\alpha = 9$, $T_\beta = 6$).

	$Len = 40, m = 40$	$Len = 20, m = 80$	$Len = 10, m = 160$
Accuracy	0.9425	0.9775	0.9875
F1 Score	0.9426	0.9772	0.9874
Sensitivity	0.945	0.96	0.98
BCE Loss	0.156	0.1499	0.1694

The VARCH method also involves two other performance-affecting parameters: T_α and T_β. After constructing Mat(B), we must determine this pair, which specifies the pairs for constructing the frequency matrix. We tested all combinations of T_α and T_β values from 4 to 14. The experiments involved converting the VARCH images into tensors, applying HDMR to these tensors, and using feature matrices with an SVM algorithm for classification accuracy. The results of these parameter trials and their impact on classification performance are shown in Table 4.2,

After testing all possible T_α and T_β pairs with short sequence lengths. We highlight the best results for each value for $T_\alpha = \{5, .., 14\}$. We determined that the optimal pair was $T_\alpha = 9$ and $T_\beta = 6$, which has the underlined results of 87.5 value. This means that once matrix A is constructed, the element in the 10-th row is selected as the x-axis, and the element in the 7-th column is selected as the y-axis. By doing this, we can also gain insights into the distribution of the best T_α and T_β pairs. The top performances,

Table 4.2 : Classification accuracy performances of T_α and T_β pairs ($Len = 20, m = 20$).

T_α/T_β	5	6	7	8	9	10	11	12	13
5	80	87	81	86	80	82	83	84	81
6	85	82	80	80	86	84	81	86	85
7	79	81	82	84	81	77	80	80	82
8	83	84	82	81	84	82	82	82	85
9	81	**87.5**	82	85	82	82	80	82	82
10	77	84	75	82	80	80	82	78	76
11	80.5	84.5	77	80	81	80	82	82	80.25
12	81.2	86.2	80	84.75	82	81.75	83.7	81	79.2
13	78.5	87.25	82	84	80.7	77	81.7	79	81.5

both for α and β, are observed for the values within the range of $\{5,..,14\}$. Table 4.2 shows that these results indicate that the combinations of two bases, specifically $N_C + N_T$ and $N_A + N_G$ in our scenario, contain more information than single three-base and four-base combinations.

Table 4.3 : VARCH ($Len = 10$, $m = 160$) and CGR ($n = 5$) results of 400 patient/control data.

	CGR	**VARCH**
Accuracy	0.97	0.9875
F1 Score	0.9701	0.9874
Recall	0.95	0.98
BCE Loss	0.69	0.1694

Based on the parameters obtained from these experiments, comparisons of CGR and VARCH methods were made. Initially, a balanced dataset from 400 patients and 400 controls was used. For the CGR method, the parameter $n = 5$ was chosen, and for the VARCH method, the parameters $T_\alpha = 9$ and $T_\beta = 6$, $Len = 10$, and $m = 160$ were selected. The HDMR method was not re-explained as it was applied similarly to the tensors obtained from both processes.

As seen in Table 4.3, in the balanced dataset experiments, the highest accuracy value for the CGR method was 0.97, while the VARCH method achieved an accuracy of 0.9875. This indicates that the VARCH method can better determine whether a gene sequence belongs to a diseased or healthy individual 1.7%. In these studies, the most critical outcome is likely to be the false negatives, where genuinely diseased

individuals are overlooked and classified as healthy. This is identified under recall calculations. In the opposite scenarios, where healthy individuals are classified as diseased, subsequent examinations will rectify this situation. Therefore, recall calculations were also considered. Results shows recall values were higher by 0.5% in the VARCH method.

We have also calculated the loss values using binary cross-entropy. When measuring prediction accuracy in binary classification, binary cross-entropy is a loss function that compares predicted probabilities to proper labels. VARCH method's loss value is significantly lower than the CGR result such as, 0.69 while other is 0.16. Lower numbers indicate better alignment between actual labels and predictions. In spam detection and disease diagnosis applications, it helps maximize model performance by ensuring accurate and trustworthy probability estimations [33].

The accuracy performance of the linear kernel SVM did not meet our expectations with imbalanced datasets, so we also conducted experiments using the RBF kernel instead of linear.

Table 4.4 : VARCH ($Len = 10, m = 160$) and CGR ($n = 5$) results of 100 patient/400 control data.

	CGR	**VARCH**
Accuracy	0.9125	0.976
F1 Score	0.8147	0.9375
Recall	0.75	0.9
BCE Loss	0.76	0.21

As seen in Table 4.4, in the experiments with the imbalanced dataset, the highest accuracy value for the CGR method was 0.9125, while the VARCH method achieved an accuracy of 0.976. This indicates that similar to the balanced dataset results, the VARCH method is better at determining whether a gene sequence belongs to a diseased or healthy individual in the imbalanced dataset. Recall values are also higher for the VARCH method, and as expected, the BCE loss is lower than the CGR.

Although accuracy, F1 score, and sensitivity values are very close in both methods, the loss values are significantly lower in the VARCH method. While the CGR method reduced the loss to 0.76, the VARCH method managed to reduce it to 0.21 on

imbalanced dataset. This is a significant difference but on the other hand computational cost is the most crucial aspect. The more successful and accurate results obtained with the VARCH method suggest that this method should be preferred. However, the comparison of computational cost and computational complexity also needs to be evaluated to determine which method is more advantageous and practical.

The overall complexity of the CGR method is primarily determined by the `chaos_game_representation` function due to its $O(4^k + pk)$ complexity. Here, k represents the length of the $k-mer$s, which are substrings of length k within a DNA sequence, while p denotes the number of unique $k-mer$s in the input sequence, a factor that depends on the sequence's length and composition.

This complexity makes the algorithm computationally intensive for large k values. However, it remains efficient in other areas with linear or near-linear complexity. Provided k is not exceedingly large, k=5 in this study, the key term to consider is $O(pk)$, where p is the number of distinct $k-mer$s.

In contrast, the VARCH method has an $O(n + N \cdot c \cdot s + m^4)$ complexity. The `Calculate_Varch` function largely dictates this complexity due to the $O(m^4)$ term, making it computationally demanding for large m values. Thus, the computational cost increases significantly when we minimize the parameter Len and maximize m to achieve the best accuracy. This method could be enhanced in the future to maintain high accuracy performance while keeping the computational cost acceptable. From this standpoint, the CGR method is more feasible for real applications with large datasets.

5. CONCLUSION

Gene network studies aim to decode an individual's genetic composition, allowing for a better understanding of health and disease implications. Advanced techniques, such as genotyping, high-throughput sequencing, and bioinformatics, have been used to study the human genome's structure, functionality, and diversity. Gene network analysis is critical in understanding how genes interact with and regulate cellular processes. This knowledge is crucial in uncovering disease mechanisms, identifying therapeutic targets, and advancing personalized medicine. Gene networks provide a comprehensive view of biological systems, offering insights into disease complexity and aiding the development of innovative treatment strategies.

In this thesis, our primary goal was to improve our understanding of the genetic patterns associated with different phenotypes. Advanced genetic visualization tools and digh dimensional data decomposition methods were used for the gene network analysis. The CGR and VARCH methods were used to visually represent genetic sequences, enabling the identification of potential genetic signatures that differentiate between the patient and control groups. We also compared the performances of these two methods. Applying the High Dimensional Model Representation (HDMR) method further simplifies the data, allowing for more accurate and efficient classification. By combining these visual and analytical approaches and using a Support Vector Machines (SVM) algorithm for classification, we aimed to enhance the accuracy of the gene classification. This could advance genetics and contribute to the development of personalized therapies and disease diagnosis.

This thesis presents the results of an analysis using an mTOR dataset. We focused on the classification accuracy obtained through the CGR, VARCH, and HDMR methods facilitated by the SVM algorithm. Our findings showed that the CGR and VARCH methods had high accuracy in classifying genetic data, with the VARCH method outperforming the CGR method. When considering recall values, which are crucial

for identifying false negatives, the VARCH method demonstrated a better performance. We also observed significantly lower BCE loss values with the VARCH method, further highlighting its effectiveness for inaccurate predictions.

The computational cost of each method was also evaluated. Although the CGR method demonstrated computational efficiency, its complexity increased with larger $k-mer$ values, affecting its practicality for large datasets. The VARCH method has higher computational demands, especially with larger parameter values, potentially limiting its use in extensive genomic analyses. Therefore, although both methods provide promising accuracy, computational complexity, and efficiency considerations are vital for their practical implementation in variant analysis workflows.

Future gene network analysis research could focus on optimizing methods like VARCH to maintain high accuracy while reducing computational complexity. Integrating deep learning techniques to improve performance and exploring hybrid approaches that combine methods such as CGR and VARCH could also be beneficial. Furthermore, thorough benchmark studies and method validation across diverse datasets will provide a more comprehensive evaluation. Finally, although we used HDMR for dimension reduction, starting from second-order terms, further benefits could be derived from the HDMR method.

REFERENCES

[1] **Steinberg, K.**, **Beck, J.**, **Nickerson, D.**, **Garcia-Closas, M.**, **Gallagher, M.**, **Caggana, M.**, **Reid, Y.**, **Cosentino, M.**, **Ji, J.**, **Johnson, D.**, **Hayes, R.B.**, **Earley, M.**, **Lorey, F.**, **Hannon, H.**, **Khoury, M.J. and Sampson, E.** (2002). DNA banking for epidemiologic studies: a review of current practices, *Epidemiology*, *13*(3), 246–254.

[2] **Travers, A. and Muskhelishvili, G.** (2015). DNA structure and function, *FEBS J.*, *282*(12), 2279–2295.

[3] **Weng, L.**, **Macciardi, F.**, **Subramanian, A.**, **Guffanti, G.**, **Potkin, S.G.**, **Yu, Z. and Xie, X.** (2011). SNP-based pathway enrichment analysis for genome-wide association studies, *BMC Bioinformatics*, *12*(1), 99.

[4] **Uffelmann, E.**, **Huang, Q.Q.**, **Munung, N.S.**, **de Vries, J.**, **Okada, Y.**, **Martin, A.R.**, **Martin, H.C.**, **Lappalainen, T. and Posthuma, D.** (2021). Genome-wide association studies, *Nat. Rev. Methods Primers*, *1*(1).

[5] **Fang, G.**, **Wang, W.**, **Paunic, V.**, **Heydari, H.**, **Costanzo, M.**, **Liu, X.**, **Liu, X.**, **VanderSluis, B.**, **Oately, B.**, **Steinbach, M.**, **Van Ness, B.**, **Schadt, E.E.**, **Pankratz, N.D.**, **Boone, C.**, **Kumar, V. and Myers, C.L.** (2019). Discovering genetic interactions bridging pathways in genome-wide association studies, *Nat. Commun.*, *10*(1), 4274.

[6] **Cano-Gamez, E. and Trynka, G.** (2020). From GWAS to function: Using functional genomics to identify the mechanisms underlying complex diseases, *Front. Genet.*, *11*, 424.

[7] **Silberstein, M.**, **Nesbit, N.**, **Cai, J. and Lee, P.H.** (2021). Pathway analysis for genome-wide genetic variation data: Analytic principles, latest developments, and new opportunities, *J. Genet. Genomics*, *48*(3), 173–183.

[8] **Wang, M.H.**, **Cordell, H.J. and Van Steen, K.** (2019). Statistical methods for genome-wide association studies, *Semin. Cancer Biol.*, *55*, 53–60.

[9] **Mutalib, S. and Mohamed, A.** (2011). A brief survey on GWAS and ML algorithms, *2011 11th International Conference on Hybrid Intelligent Systems (HIS)*, IEEE.

[10] **Ebert, B.L. and Golub, T.R.** (2004). Genomic approaches to hematologic malignancies, *Blood*, *104*(4), 923–932.

[11] **Wang, J.**, **Kong, L.**, **Gao, G. and Luo, J.** (2013). A brief introduction to web-based genome browsers, *Brief. Bioinform.*, *14*(2), 131–143.

[12] **Rehm, B.H.** (2001). Bioinformatic tools for DNA/protein sequence analysis, functional assignment of genes and protein classification, *Appl. Microbiol. Biotechnol.*, *57*(5-6), 579–592.

[13] **Banerjee, S.**, **Zeng, L.**, **Schunkert, H. and Söding, J.** (2018). Bayesian multiple logistic regression for case-control GWAS, *PLoS Genet.*, *14*(12), e1007856.

[14] **Ik-Soo, H.**, **Sohee-Oh and Park, T.** (2011). A chi-square test for detecting multiple joint genetic variants in genome-wide association studies, *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, IEEE.

[15] **Romagnoni, A.**, **Jégou, S.**, **Van Steen, K.**, **Wainrib, G.**, **Hugot, J.P. and International Inflammatory Bowel Disease Genetics Consortium (IIBDGC)** (2019). Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data, *Sci. Rep.*, *9*(1), 10351.

[16] **Nguyen, T.T.**, **Huang, J.**, **Wu, Q.**, **Nguyen, T. and Li, M.** (2015). Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests, *BMC Genomics*, *16 Suppl 2*(S2), S5.

[17] **Botta, V.**, **Louppe, G.**, **Geurts, P. and Wehenkel, L.** (2014). Exploiting SNP correlations within random forest for genome-wide association studies, *PLoS One*, *9*(4), e93379.

[18] **Liu, Y.**, **Wang, D.**, **He, F.**, **Wang, J.**, **Joshi, T. and Xu, D.** (2019). Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean, *Front. Genet.*, *10*, 1091.

[19] **Mostavi, M.**, **Chiu, Y.C.**, **Huang, Y. and Chen, Y.** (2020). Convolutional neural network models for cancer type prediction based on gene expression, *BMC Med. Genomics*, *13*(Suppl 5), 44.

[20] **Jeffrey, H.J.** (1990). Chaos game representation of gene structure, *Nucleic Acids Res.*, *18*(8), 2163–2170.

[21] **Wu, R.**, **Liu, W.**, **Mao, Y. and Zheng, J.Z.J.** (2020). 2D graphical representation of DNA sequences based on variant map, *IEEE Access*, *8*, 173755–173765.

[22] **Rabitz, H.**, **Aliş, Ö.F.**, **Shorter, J. and Shim, K.** (1999). Efficient input–output model representations, *Comput. Phys. Commun.*, *117*(1-2), 11–20.

[23] **Barabási, A.L.**, **Gulbahce, N. and Loscalzo, J.** (2011). Network medicine: a network-based approach to human disease, *Nat. Rev. Genet.*, *12*(1), 56–68.

[24] **Emilsson, V.**, **Thorleifsson, G.**, **Zhang, B.**, **Leonardson, A.S.**, **Zink, F.**, **Zhu, J.**, **Carlson, S.**, **Helgason, A.**, **Walters, G.B.**, **Gunnarsdottir, S.**, **Mouy, M.**, **Steinthorsdottir, V.**, **Eiriksdottir** *et al.* (2008). Genetics of gene expression and its effect on disease, *Nature*, *452*(7186), 423–428.

[25] **Carninci, P.**, **Kasukawa, T.**, **Katayama, S.**, **Gough, J.**, **Frith, M.C.**, **Maeda, N.**, **Oyama, R.**, **Ravasi, T.**, **Lenhard, B.**, **Wells, C.**, **Kodzius, R.**, **Shimokawa, K.**, **Bajic, V.B.**, **Brenner, S.E.**, **Batalov, S.**, **Forrest, A.R.R.**, **Zavolan, M.**, **Davis, M.J.**, **Wilming** *et al.* (2005). The transcriptional landscape of the mammalian genome, *Science*, *309*(5740), 1559–1563.

[26] **Lewis, B.P.**, **Burge, C.B. and Bartel, D.P.** (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets, *Cell*, *120*(1), 15–20.

[27] **Wang, Y.X.R. and Huang, H.** (2014). Review on statistical methods for gene network reconstruction using expression data, *J. Theor. Biol.*, *362*, 53–61.

[28] **Xie, X.**, **Kendzior, M.C.**, **Ge, X.**, **Mainzer, L.S. and Sinha, S.** (2021). VarSAn: associating pathways with a set of genomic variants using network analysis, *Nucleic Acids Res.*, *49*(15), 8471–8487.

[29] **Konuma, T. and Okada, Y.** (2021). Statistical genetics and polygenic risk score for precision medicine, *Inflamm. Regen.*, *41*(1), 18.

[30] **Lewis, C.M. and Vassos, E.** (2020). Polygenic risk scores: from research tools to clinical instruments, *Genome Med.*, *12*(1), 44.

[31] **Vilhjálmsson, B.J.**, **Yang, J.**, **Finucane, H.K.**, **Gusev, A.**, **Lindström, S.**, **Ripke, u.**, **Loh, P.R.**, **Bhatia, G.**, **Do, R.**, **Hayeck, S.**, **Genovese, G.** *et al.* (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores, *97*(4), 576–592.

[32] **Lloyd-Jones, L.R.**, **Zeng, J.**, **Sidorenko, J.**, **Yengo, L.**, **Moser, G.**, **Kemper, K.E.**, **Wang, H.**, **Zheng, Z.**, **Magi, R.**, **Esko, T.**, **Metspalu, A.**, **Wray, N.R.**, **Goddard, M.E.**, **Yang, J. and Visscher, P.M.** (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics, *Nat. Commun.*, *10*(1), 5086.

[33] **Anderson, C.A.**, **Pettersson, F.H.**, **Clarke, G.M.**, **Cardon, L.R.**, **Morris, A.P. and Zondervan, K.T.** (2010). Data quality control in genetic case-control association studies, *Nat. Protoc.*, *5*(9), 1564–1573.

[34] **Price, A.L.**, **Patterson, N.J.**, **Plenge, R.M.**, **Weinblatt, M.E.**, **Shadick, N.A. and Reich, D.** (2006). Principal components analysis corrects for stratification in genome-wide association studies, *Nat. Genet.*, *38*(8), 904–909.

[35] **Aschard, H.**, **Hancock, D.B.**, **London, S.J. and Kraft, P.** (2010). Genome-wide meta-analysis of joint tests for genetic and gene-environment interaction effects, *Hum. Hered.*, *70*(4), 292–300.

[36] **Marchini, J. and Howie, B.** (2010). Genotype imputation for genome-wide association studies, *Nat. Rev. Genet.*, *11*(7), 499–511.

[37] **Laird, N.M. and Mosteller, F.** (1990). Some statistical methods for combining experimental results, *Int. J. Technol. Assess. Health Care*, *6*(1), 5–30.

[38] **Almeida, J.S.**, **Carriço, J.A.**, **Maretzek, A.**, **Noble, P.A. and Fletcher, M.** (2001). Analysis of genomic sequences by Chaos Game Representation, *Bioinformatics*, *17*(5), 429–437.

[39] **Rabitz, H. and Aliş, Ö.F.** (1999). General foundations of high-dimensional model representations, *J. Math. Chem.*, *25*(2/3), 197–233.

[40] **Tuna, S.**, **Gulec, C.**, **Yucesan, E.**, **Cirakoglu, A. and Arguden, Y.T.** (2023). Gene teams are on the field: Evaluation of variants in gene-networks using high dimensional modelling, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, *20*(5), 2959–2969.

[41] **Hoang, T.**, **Yin, C. and Yau, S.S.T.** (2016). Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison, *Genomics*, *108*(3-4), 134–142.

[42] **Zheng, J.Z.J. and Zheng, C.H.** (2010). A framework to express variant and invariant functional spaces for binary logic, *Front. Electr. Electron. Eng. China*, *5*(2), 163–172.

[43] **Taylor, B.**, **Lee, T.J. and Weitz, J.S.** (2013). A guide to sensitivity analysis of quantitative models of gene expression dynamics, *Methods*, *62*(1), 109–120.

[44] **Li, G. and Rabitz, H.** (2012). General formulation of HDMR component functions with independent and correlated variables, *J. Math. Chem.*, *50*(1), 99–130.

[45] **Chapelle, O.**, **Haffner, P. and Vapnik, V.N.** (1999). Support vector machines for histogram-based image classification, *IEEE Trans. Neural Netw.*, *10*(5), 1055–1064.

[46] **Dagher, I.** (2008). Quadratic kernel-free non-linear support vector machine, *J. Glob. Optim.*, *41*(1), 15–30.

[47] **Wullschleger, S.**, **Loewith, R. and Hall, M.N.** (2006). TOR signaling in growth and metabolism, *Cell*, *124*(3), 471–484.

[48] **Saxton, R.A. and Sabatini, D.M.** (2017). MTOR signaling in growth, metabolism, and disease, *Cell*, *168*(6), 960–976.

[49] **Laplante, M. and Sabatini, D.M.** (2012). mTOR signaling in growth control and disease, *Cell*, *149*(2), 274–293.

[50] **Guertin, D.A. and Sabatini, D.M.** (2005). An expanding role for mTOR in cancer, *Trends Mol. Med.*, *11*(8), 353–361.

[51] **Perl, A.** (2015). mTOR activation is a biomarker and a central pathway to autoimmune disorders, cancer, obesity, and aging, *Ann. N. Y. Acad. Sci.*, *1346*(1), 33–44.

[52] **Szwed, A.**, **Kim, E. and Jacinto, E.** (2021). Regulation and metabolic functions of mTORC1 and mTORC2, *Physiol. Rev.*, *101*(3), 1371–1426.

CURRICULUM VITAE

Name SURNAME: Pınar GÜLER

EDUCATION:

- **B.Sc.:** 2020, Istanbul Technical University, Faculty of Electrical and Electronics Engineering, Control and Automation Engineering

PROFESSIONAL EXPERIENCE AND REWARDS:

- 2021-2024 Siemens AG, Software Developer
- 2020-2021 Patika Robotics, Robotics Software Engineer

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

- Güler P., Tuna S. (2024), Visualization Based Analysis of Gene Networks Using High Dimensional Model Representation, 3rd International Graduate Research Symposium (IGRS'24) May 8-10, 2024 Istanbul, Türkiye. (Oral Presentation)