



**NEAR EAST UNIVERSITY
INSTITUTE OF GRADUATE STUDIES
DEPARTMENT OF BIOMEDICAL ENGINEERING**

**THE IMPACT OF ENCODING METHODS ON sgRNA
PREDICTIONS FOR CRISPR-CAS12 IN ENHANCING
GENOME EDITING**

M.Sc. THESIS

EYAD AHMAD MUSTAFA AL ZOUBI

**Nicosia
September, 2023**

EYAD AL ZOUBI

**THE IMPACT OF
ENCODING METHODS
ON sgRNA
PREDICTIONS FOR**

MASTER THESIS

**NEU
2023**

**NEAR EAST UNIVERSITY
INSTITUTE OF GRADUATE STUDIES
DEPARTMENT OF BIOMEDICAL ENGINEERING**

**THE IMPACT OF ENCODING METHODS ON sgRNA
PREDICTIONS FOR CRISPR-CAS12 IN ENHANCING
GENOME EDITING**

M.Sc. THESIS

EYAD AHMAD MUSTAFAALZOUBI

Supervisor

Asst. Prof Zubaida Sa'id Ameen

Nicosia

September, 2021

Approval

We certify that we have read the thesis submitted by EYAD AHMAD titled "THE IMPACT OF ENCODING METHODS ON *RNA PREDICTIONS FOR CRISPRCAS12 IN ENHANCINGGENOME EDITING" and that in our combined opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Educational Sciences.

Examining Committee	Name-Surname	Signature
---------------------	--------------	-----------

Head of the Committee: Assoc. Prof. Dr. Suleyman Asir

Committee Member* : Asst. Prof .Dr. Abdullahi Garba Usman

Committee Member* : Dr .Abubakar Adamu

Supervisor :Asst. Prof .Dr.Zubaida Sa'id Ameen

Co-supervisor:Asst. Prof, Dr. Auwalu Saleh Mubarak



Approved by the Head of the Department


16/1/2024



Prof. Dr. Suleyman Asir
Head of Department

Approved by the Institute of Graduate Studies

16/1/2024



Prof. Dr. Kemal Hüsnü Can Başer
Head of the Institute



Declaration

I hereby declare that all information, documents, analysis and results in this thesis have been collected and presented according to the academic rules and ethical guidelines of Institute of Graduate Studies, Near East University. I also declare that as required by these rules and conduct, I have fully cited and referenced information and data that are not original to this study.

Name and Surname of the Student

...../...../.....

Day/Month/Year

Acknowledgments

I want to sincerely thank Asst. Prof Zubaida Said Ameen, my thesis advisor, for all of your help, encouragement, and support during my research. I am also appreciative of Near East University for giving me the tools I needed to do my research, and I thank the members of my thesis committee for their insightful remarks and recommendations. I want to express my gratitude to my family, friends, and fiancée for their continuous encouragement and support over my academic career. Lastly, I want to express my gratitude to God for enabling me to overcome all of the challenges. Day by day, I have felt your guidance. It is you who gave me permission to complete my degree. I will continue to have faith in you for the future.

EYAD AHMAD MUSTAFA

Abstract

THE IMPACT OF ENCODING METHODS ON sgRNA PREDICTIONS FOR CRISPR-CAS12 IN ENHANCING GENOME EDITING

EYAD AHMAD MUSTAFA AL ZOUBI

MA/PhD, Department of Biomedical Engineering

June, 2023, (65) pages

CRISPR/Cas systems are incredibly robust and have the capacity to alter whole genomes. One of the most important challenges among the possible problems with CRISPR/Cas systems is developing sgRNA. Double-stranded DNA was thought to be separate by the Cas enzyme-sgRNA combination when it located a target sequence that matched. However, several sgRNAs turned out to be either inert or ineffectual. Therefore, before using a set of sgRNAs in genome editing research, it is imperative to confirm their effectiveness, which makes improving sgRNA design a worthwhile goal. Since most of the tools available are primarily made for the CRISPR/Cas9 system, this study introduced a number of intelligent machine learning models to predict the activity of sgRNA for the CRISPR/Cas12a system.

For the prediction of CRISPR/Cas12a sgRNA activity, different encoding techniques One-Hot, K-mers and Integer encoding effects were evaluated based on the performance achieved by four different machine-learning models namely, Support Vector Regressor (SVR), Random Forest (RF), Decision Tree (DT) and XGBoost. Next, to demonstrate the effect of data scaling after each encoding was used to further improve the performance of the models.

The goal was to combine the unique properties of each model to create an excellent model for CRISPR/Cas12a sgRNA activity prediction. When compared to state-of-the-art models, the findings showed exceptional performance. This thesis will help with the design and selection of active sgRNA for genome editing with the CRISPR/Cas12a system.

Keywords: CRISPR/Cas12, sgRNA activity; Cas enzyme; Machine Learning, Encoding.

Soyut

GENOM DÜZENLEMESİNİ GELİŞTİRMEDE CRISPR-CAS12 İÇİN KODLAMA YÖNTEMLERİNİN sgRNA TAHMİNLERİ ÜZERİNDEKİ ETKİSİ

EYAD AHMAD MUSTAFA AL ZOUBI

Yüksek Lisans/Doktora, Biyomedikal Mühendisliği Bölümü

Jane,2023 , (65) sayfa

CRISPR/Cas sistemleri inanılmaz derecede sağlamdır ve tüm genomu değiştirme kapasitesine sahiptir. CRISPR/Cas sistemlerinde yaşanabilecek olası problemler arasında en önemli zorluklardan biri sgRNA'nın geliştirilmesidir. Çift sarmallı DNA'nın, eşleşen bir hedef dizi bulunduğunda Cas enzimi-sgRNA kombinasyonu ile ayrı olduğu düşünülüyordu. Bununla birlikte, birçok sgRNA'nın ya inert ya da etkisiz olduğu ortaya çıktı. Bu nedenle, genom düzenleme araştırmalarında bir dizi sgRNA'yı kullanmadan önce, bunların etkililiğini doğrulamak zorunludur; bu da sgRNA tasarımını iyileştirmeyi değerli bir hedef haline getirir. Mevcut araçların çoğu öncelikle CRISPR/Cas9 sistemi için yapıldığından, bu çalışma, CRISPR/Cas12a sistemi için .sgRNA'nın aktivitesini tahmin etmek amacıyla bir dizi akıllı makine öğrenimi modeli sunmuştur

CRISPR/Cas12a sgRNA aktivitesinin tahmini için farklı kodlama teknikleri One-Hot, K-mers ve Integer kodlama etkileri, Support Vector Regressor (SVR), Random Forest(RF) olmak üzere dört farklı makine öğrenme modelinin elde ettiği performansa dayalı olarak değerlendirildi.), Karar Ağacı(DT) ve XGBoost. Daha sonra, her kodlamadan sonra veri ölçeklemenin etkisini göstermek için modellerin performansını daha da artırmak amacıyla kullanıldı

Amaç, CRISPR/Cas12a sgRNA aktivite tahmini için mükemmel bir model oluşturmak üzere her modelin benzersiz özelliklerini birleştirmektir. Son teknoloji modellerle karşılaştırıldığında bulgular olağanüstü performans gösterdi. Bu tez, CRISPR/Cas12a sistemi ile genom düzenleme için aktif sgRNA'nın tasarımı ve seçimine yardımcı olacaktır

Anahtar Kelimeler: CRISPR/Cas12, sgRNA aktivitesi; Cas enzimi; Makine Öğrenimi, Kodlama.

Table of Contents

Approval	1
Declaration	2
Acknowledgements	3
Abstract	4
Summary	5
Table of Contents	6
List of Tables/ List of Figures.....	7
List of Abbreviations.....	8

CHAPTER I

Introduction.....	9
1.1 Background and context of CRISPR-Cas technolog.....	13
1.1.1 Explanation of the CRISPR-Cas system and its revolutionary impact on genome editing.....	13
1.1.2 Historical development and key milestones in the field.....	14
1.2 Importance and relevance of genome engineering in the context of COVID19...15	
1.2.1 Overview of the COVID-19 pandemic and the need for advanced genetic	
Tools	15
1.2.2 Discussion on the potential applications of CRISPR-Cas12 in addressing COVID-19.	15
1.3 Research Problem and Statement.....	16
1.4 Research Questions / Hypotheses	19
1.5 Thesis Organization	22

CHAPTER II

2.1 Overview of CRISPR-Cas and its applications.....	24
2.2 Detailed explanation of the CRISPR-Cas system.....	26
2.3 Discussion on its various applications beyond genome editing	27
2.4 Previous studies on CRISPR-Cas in genome engineering.....	27
2.5 Review of literature related to COVID-19 and genome editing.....	27
2.6 Review of literature related to COVID-19 and genome editing.....	28
2.7 Deep Learning in Genomics.....	28
2.7.1 Introduction to deep learning techniques and their applications in genomics.....	29
2.7.2 Review of studies using deep learning for analysing genetic data.....	31
2.8 Examination of indel frequencies and their implications	32
2.8.1 Explanation of indel frequencies and their significance in genome editing	33
2.8.2 Review of studies exploring indel frequencies in various contexts	33
2.9 NANOTECHNOLOGY AND Artificial Intelligence/Machine learning	33

CHAPTER III

Predictive Data Modelling of Crispr-cas12 Using Various Encoding Methodologies...34	
3.1 Introduction	34
3.2 The Dataset Description.....	34
3.3 The Encoding Techniques.....	35
3.4 Results and Discussion.....	35
3.5 Conclusion.....	43

CHAPTER IV

CAS12 Crispr Data: Prediction With Various DNA Sequence Encodings And Data Scaling.....	44
4.1 Introduction	44
4.2 The Dataset Description.....	44
4.3 The Encoding Technique.....	46
4.4 Results And Discussion.....	48
4.5 Conclusion.....	49
CHAPTER V	
Conclusion and Recommendations	69
CHAPTER VI	
REFERENCES	73

List of Tables

Table 3.1 K-mers encoding	39
Table 3.2 One-hot encoding	40
Table 3.3 Integer encoding	40

List of Figures

Figure 4.1 data set description	44
Figure 5.1 visualization of nucleotide frequency	48

List of Abbreviations

Proteins Utilized for Genome Engineering: PUGE

C12:	CRISPR-Cas12
C19:	Coronavirus Disease 2019 (COVID-19)
IF :	Indel Frequency
ML:	Machine Learning
AI:	Artificial intelligence
CRISPR:	Clustered Regularly Interspaced Short Palindromic Repeats
crRNA :	CRISPR RNA
DSB :	Double-Strand Break
CNN :	Convolutional neural network
NHEJ :	Non-Homologous End-Joining
PAM :	Protospacer Adjacent Motif
DNA :	Deoxyribonucleic acid
sgRNA :	Single Guide RNA
TALEN :	Transcription Activator-like effector nuclease
ZFN :	Zinc Finger Nuclease
Tra-crRNA :	trans-activating CRISPR RNA
RNP:	ribonucleoproteins

ssDNA:	Single-Stranded DNA
siRNA:	Short interfering ribonucleic acid
LOD :	Limit of detection
LOQ :	Limit of Quantification
LSTM :	Long short-term memory
HDR:	homology-directed repair
ML:	Machine learning
DL :	Deep Learning
Acr :	anti-CRISPR
AAV :	Adeno-associated virus
gRNA :	guide RNA
NT :	Nucleotide
Ds DNA :	double-stranded DNA
Cpf1 :	CRISPR-associated endonuclease in <i>Prevotella</i> and <i>Francisella</i>
Pre-crRNA :	precursor CRISPR RNA
dCas9 :	dead Cas9

CHAPTER I

Introduction

Covid-19 is a terrible disease that affects millions of people around the world. The complexity and heterogeneity of this disease mean that despite improvements in medical research and treatment options, COVID remains a serious health risk. Traditional strategies to detect and treat it often fail to address the many genetic changes that contribute to the emergence and progression of COVID-19(Cascella et al., 2022).

On the other hand, recent developments in CRISPR cas9 technologies and artificial intelligence (AI) are opening new horizons for improving the care of Covid-19 patients. Artificial intelligence has developed into a powerful tool in the medical profession, especially in the medical field(Esteva et al., 2019).

The ability to examine large data sets and extract important insights has transformed disease detection and treatment. AI algorithms may analyze various data types, including genetic data, medical records, and imaging data, to uncover patterns and correlations that human analysis alone may ignore. Using artificial intelligence, researchers can build predictive models to assess risk, early diagnose diseases, and predict response to treatment.

CRISPR (clustered regularly interspaced short palindromic repeats) technology has revolutionized genetic engineering. Due to their precision and versatility in modifying DNA sequences, targeted therapies now have additional possibilities. Using CRISPR, researchers may be able to fix disease-causing mutations or increase the effectiveness of existing treatments. CRISPR technology, which allows the editing of genes associated with the onset of disease, holds promise for personalized treatment. The combination of artificial intelligence and CRISPR technology has a lot to offer in the field of disease detection and treatment. Complex genomic data can be examined by AI-based algorithms to detect genetic mutations and predict the risk of developing COVID-19. Using CRISPR and this understanding, programs can be developed using precise gene editing to remove or correct mutations that cause COVID-19. This method may contribute to the creation of more efficient drugs that are customized according to the specific genetic profile of each patient, thus increasing treatment success rates and reducing adverse effects. While using AI and CRISPR to treat the disease offers exciting new prospects, it also raises many challenges. It is crucial to comprehensively evaluate the ethical implications of using CRISPR and AI in humans. Important considerations

include appropriate and open use of patient data, informed permission, and protection of privacy. Furthermore, legislative frameworks must be built to ensure the ethical and moral use of CRISPR and AI technologies in clinical practice(Y. Zhang et al., 2021).

Finally, the combination of AI and CRISPR technology offers an exciting new frontier in Covid-19 detection and treatment. AI algorithms combined with genome analysis and CRISPR-enabled gene editing have the potential to transform Covid-19 treatment by allowing earlier detection, personalized treatment, and improved patient outcomes. However, further study, clinical trials, and collaboration across multidisciplinary teams are needed to properly explore the revolutionary potential of this strategy.

1.1 Background and context of CRISPR-Cas technology

AI systems rely on large amounts of diverse, high-quality data to create accurate predictions and diagnoses. In the context of disease diagnosis, combining data from multiple sources such as genetics, imaging, electronic health records, and clinical trials is crucial. The challenge is to collect and evaluate this disparate data to obtain relevant insights that can help accurately diagnose diseases and plan treatment.

Covid-19 is a very diverse disease, with each patient having distinct characteristics and responses to treatment. Precision and personalized medicine are made possible by artificial intelligence and CRISPR technology, which tailors treatment to patients. However, using CRISPR to build algorithms and methodologies that can rapidly evaluate data from individual patients, detect genetic changes, predict treatment response and guide focused treatment is a challenging task.

The use of AI and CRISPR technologies to detect and treat Covid-19 raises ethical and legal concerns. One major issue is requiring AI algorithms to make open and responsible decisions, which is essential given privacy concerns over the use of patient data and the risks associated with off-target consequences of CRISPR gene editing. It is crucial to create strong ethical and governance frameworks that address these challenges while continuing to encourage innovation.

Although CRISPR and artificial intelligence have shown great promise in research settings, translating these technologies into practical clinical applications is difficult. To apply AI algorithms to real-time diagnosis of communicable diseases and make treatment decisions

requires comprehensive validation, integration with existing healthcare systems, and overcoming professional acceptance hurdles. Additionally, for these technologies to be successfully integrated into clinical operations, healthcare practitioners must be educated and appropriately trained in their use.

Studying the use of artificial intelligence (AI) and CRISPR technology in diagnosing and treating Covid-19 has limitations. Obtaining diverse, high-quality datasets to train AI algorithms is difficult due to data availability and privacy concerns. AI algorithms rely on data patterns, which may not fully capture the biological mechanisms of COVID-19. Integrating AI and CRISPR into clinical practice requires validation, regulatory approvals, and ethical considerations, which may slow its implementation. In addition, the potential off-target effects and long-term safety of CRISPR gene editing must be carefully evaluated. Overcoming these limitations is essential to responsibly and effectively use AI and CRISPR in COVID-19 patient care. To achieve our goal, we must overcome this barrier by gaining better knowledge of their mistakes and strategies to eliminate or reduce them in the coming years.

In this section of the report, we will discuss the most important terms mentioned. This will allow the reader to get a better understanding of the entire topic.

- **Artificial Intelligence (AI):** Related to the development of computer systems and algorithms capable of performing activities that normally require human intelligence. In the context of diagnosing and treating Covid-19, artificial intelligence (AI) is being used to evaluate and understand vast amounts of data, such as genetic data, medical imaging and clinical records, to aid accurate diagnosis, plan treatment and predict outcomes.
- **CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats):**

CRISPR is an innovative gene editing method that allows scientists to precisely change DNA sequences. It is based on a mechanism found in bacteria that uses the CRISPR-associated Cas protein to target and edit specific regions in DNA. In the case of COVID-19, CRISPR technology could be used to target and modify genes associated with the disease, which could open up new treatment options by preventing genetic mutations that lead to disease development or improving the immune system's ability to identify and destroy disease-causing cells.

1.1.1 Overview of CRISPR-Cas9 and cas 12 system and its revolutionary impact on genome editing.

The CRISPR-Cas framework is made up of the CRISPR-associated Cas protein and its short, frequently spaced recombination events. This framework has revolutionized research in the life sciences by providing numerous tools for modifying, identifying, capturing, and annotating clear data. The design of various natural substances' DNA or RNA. In this method, foreign DNA extracts, also known as spacers, are inserted into CRISPR cassettes, converted into CRISPR mappings, and processed to produce guide RNA-gRNA. The characteristics of Cas proteins shape CRISPR designs. In the hunt for novel transgressive-targeting spacers, CAS proteins present a potential enzymatic tool. Some Cas proteins, including all-cases, have been used to create novel genome design tools due to the specificity of their programming combinations. (Hillary & Ceasar, 2023).

A new class of CRISPR genome editing tools that include primer editors and DNA base editors. The two most fundamental methods for repairing these fractures are homology-directed repair (HDR) and non-homologous end-joining repair (NHEJ) bypasses the main flaws of CRISPR Cas9 systems, such as a limited editing efficiency and a small risk of off-target consequences. (Kass & Jasin, 2014)

Researchers from the Broad Institute of MIT and Harvard University identified and characterized the Cas12a system, which is a type V member of CRISPR nucleases found in *Prevotella* and *Francisella* 1 bacteria. (Zetsche et al., 2015).

There are various ways in which CRISPR-Cas12a differs from Cas9. A tracrRNA is not necessary for the single crRNA nuclease CRISPR-Cas12a to function. Target DNA is cleaved by CRISPR-Cas12a distal to the PAM site, producing cohesive rather than blunt ends and 4- or 5-nt overhangs. In contrast to the Cas9 system, which uses PAM "N₃GG," Cas12a requires the PAM sequence "TTN/TTTN/TTTV," where N= A/T/C/G and V=A/C/G. (Fonfara et al., 2016). Because of its special qualities, CRISPR-Cas12a is a useful addition to the toolkit for CRISPR-based genome editing, allowing for a greater variety of targetable genomic regions. However, the development and deployment of the CRISPR/Cas system is impeded by sgRNA (on-target) activity. (Xu et al., 2015).

1.1.2 Historical development and key milestones in the field.

The exact altering of DNA groupings inside living cells made conceivable by the 2012 improvement of CRISPR and Cas9 innovation fundamentally affected science. In 1970, genetically engineered mice were used to begin genome editing. (Yoshida et al., 2022).

The CRISPR Cas9 technology is changing fields like genome editing, which is changing science and technology. It builds on groundbreaking discoveries like the structure of double-stranded DNA, in vitro fertilization, and the cloning of Dolly the sheep. Mainstream researchers will watch out for guidelines, moral standards, and potential purposes of genome altering in different spaces. (Gostimskaya, 2022).

The mid-1990s saw the discovery of an unexpectedly unusual sequence in the Escherichia coli genome. CRISPR sequences and Cas proteins, which are now thought to be extremely powerful adaptive immunity mechanisms, are found in prokaryotes. Modern genetic engineering techniques have been made possible by the discovery of Cas CRISPR, a significant advancement in fundamental biology. However, the discovery of CRISPR activities has been made feasible by the genomics revolution and bioinformatics technologies. (Ishino et al., 2018).

Genetic information is made up of deoxyribonucleic acid (DNA) and genes. Genes encode proteins and RNA and provide instructions for living things' growth, operation, and reproduction. Because nucleotide sequences can be altered by both internal and external stimuli, resulting in mutations that impede cellular activities and cause diseases like cancer, the idea of restoring original DNA sequences as a treatment has been floated. (Matsumoto & Nomura, 2023).

1.2 Importance and relevance of genome engineering in the context of COVID-19

The biology, variety, and development of the SARS-CoV-2 virus, which has had a significant impact on the COVID-19 pandemic, can now be studied because of advances in genomics. As a result, a new era of genome monitoring has begun, with the possibility of real-time monitoring of genome changes. Modern vaccines are made possible by high-throughput sequencing and gene editing technologies, and the availability of genomic and proteomic data makes it easier to identify and treat molecular diseases. (Ameen et al., 2021).

1.2.1 Overview of the COVID-19 pandemic and the need for advanced genetic tools.

We'll go over the two distinct CRISPR-Cas systems and the functions of each Cas protein, including the three Cas proteins now known as Cas12, 13, and 14. In addition, we will discuss the benefits and drawbacks of the novel application of Cas protein in several domains, such as the detection of coronaviruses that are responsible for severe acute respiratory illnesses. ~~19~~ Next-generation precision editing of the SARS-CoV genome may benefit from the functional properties of various Cas proteins. (Hillary & Ceasar, 2023)

1.2.2 Discussion on the potential applications of CRISPR-Cas12

Cas12a may be helpful in genome editing because of its capacity to promote particular integration as a result of staggered incisions. Soybeans are among the plants whose genomes have been altered through the use of Cas12a. (Kim et al., 2017). Ag development can be accelerated greatly through genome engineering. One of the first institutions to create legislation pertaining to genome engineering was the European Union. (Friedrichs et al., 2019). Three categories were established for genome editing. Small indels or single nucleotide alterations are the result of non-homologous end-joining-induced site-directed nuclease (SDN) 1 events. SDN2 events alter a few nucleotides by employing a template. For instance, template-mediated repair was utilized to produce herbicide-resistant mutant strains of rice. Using a repair template containing the point mutations of interest, LbCas12a was utilized to produce staggered breaks in the Acetolactate synthase (ALS) gene, and the HDR approach was employed to precisely insert small nucleotide alterations. (Li et al., 2018). The most stringent evaluations are probably going to come from inserting foreign DNA from another species (SDN3), which is almost certainly going to be categorized as transgenic in most countries. The United States Department of Agriculture (USDA), the Food and Drug Administration (FDA), and the Environmental Protection Agency (EPA) are all involved in the regulatory process regarding modified plant products in the US. Therefore, it is still challenging to forecast when and how much it will cost to introduce genome-edited products to the market. (Schmidt et al., 2020).

1.3 Research Problem and Statement

Artificial intelligence (AI) in COVID-19 detection and treatment using CRISPR is being researched to fully realize the potential of these innovative technologies to transform Covid-19 patient care. AI can evaluate multiple data sets from multiple sources, enabling accurate COVID diagnosis and personalized treatment guidance. Given its precise gene-editing capabilities, CRISPR holds promise for targeted drugs and overcoming genetic abnormalities that underpin Covid-19 development. By combining the analytical skills of artificial intelligence with the precision of CRISPR, researchers expect to increase early diagnosis, improve treatment outcomes, and develop innovative medicines for different types of Covid disease. The goal is to improve the quality of life of Covid-19 patients and save lives by enhancing our understanding of the biology of the disease and transforming this understanding into effective clinical interventions.

The main mechanism by which the CRISPR/Cas system modifies genomes is sgRNA migration. While indel frequencies produced by distinct sgRNAs vary, indel recurrences generated subsequent to CRISPR exploration can provide information about sgRNA mobility. The fact that different sgRNAs function differently and produce unequal activity is still a major issue (Moreno-Mateos et al., 2015). The three forms of current methods are alignment, hypothesis, and machine learning approaches; using machine learning is the suggested strategy, though (Yan et al., 2018).

Artificial intelligence methods predict sgRNA cleavage activity by cultivating a model that takes into account multiple factors that impact action (Van Der Oost et al., 2014). Furthermore, manually created characteristics could produce redundant data and subpar prediction results. Consequently, there are definite limitations to AI-based methods, such as the need for low speculation and master topic knowledge. Deep learning understanding (Lecun et al., 2015).

The primary goal of the study is to predict CRISPR/Cas12a sgRNA activity (Indel frequency) using deep learning techniques.

The objectives of this study are:

- To evaluate different sequence encoding techniques for the prediction of Indel Frequencies
- To compare different models to performance in predicting the indel frequency

- To show the effect of scaling after employing different sequence encoding
Creating various models with various hyperparameters and evaluating how well they work.
- To assess how well deep learning models predict sgRNA activity in comparison to alternative machine learning algorithms.
- To comprehend the characteristics that go into predicting sgRNA activity.
- To forecast sgRNA activity in order to detect COVID-19.
- To apply transfer learning to small dataset prediction tasks.

1.4 Research Questions / Hypotheses

The primary goal of the study is to forecast CRISPR/Cas12a sgRNA activity (Indel frequency) using machine learning techniques.

The objectives of this study are:

- Will different sequence encoding techniques affect the prediction of Indel Frequencies
- What will be the effect of scaling after employing different sequence encoding
- What will be the performance of different models with various hyperparameters and assess how well they work

1.5 Thesis Organization

The thesis report is divided into five chapters, each of which completely explains the research aims. Chapters 3 and 4 can be read separately or together. The chapters are listed as follows:

Chapter1: This chapter provides the background information for the study. This chapter also covers the problem statement and objectives of the study.

Chapter 2: A review of anti-CRISP and the Cas12 system's applications. A thorough and succinct explanation of the techniques for forecasting sgRNA activity is also included.

Chapter 3: This chapter shows PREDICTIVE DATA MODELING OF CRISPR-CAS12 USING VARIOUS ENCODING METHODOLOGIES.

Chapter 4: This chapter detailed CAS12 CRISPR DATA: PREDICTION WITH VARIOUS DNA SEQUENCE ENCODINGS AND DATA SCALING.

Chapter 5: This chapter contains the thesis work's conclusion as well as suggestions for more research.

CHAPTER II

Literature Review

2.1 Overview of CRISPR-Cas and its applications.

The CRISPR-Cas9 system enables prokaryotes to naturally protect themselves from viruses by detecting and filtering exogenous genomic elements. It is comprised of an aide RNA and the Cas9 protein, and it is incorporated into prokaryotes through procured insusceptibility. There are two steps to this strategy: creating guide RNA and determining the target gene This system has numerous uses in molecular biology, ranging from fundamental research to practical applications. Although significant progress has been made, practical implementations continue to face difficulties. Improvements are needed to maximize benefits while minimizing risks.(Y. Zhang et al., 2021). Originally designed as a bacterial defense mechanism against phages and other transportable genetic elements like plasmids and transposons, the CRISPR-Cas architecture(Hille et al., 2018).

Three significant developments are involved in the enhancement of CRISPR-Cas frameworks:

(i) CRISPR transformation, which entails inserting foreign, attacking genetic segments into a CRISPR display as spacer successions.

(ii) CrRNA development: The CRISPR exhibit is translated into pre-crRNA, which is then modified into mature crRNAs. At that juncture, the mature crRNAs use Cas effector proteins to construct crRNA effector buildings.

- (i) (iii) CRISPR impedance - By promoting grouping explicit obliteration, these crRNA-Cas structures identify and eliminate foreign genomic segments (Jackson et al., 2017).

The study of artificial intelligence (AI) in covid 19 diagnosis and treatment using CRISPR is of great importance. It has the potential to revolutionize the care of COVID-19 patients by enhancing early detection, improving accuracy in diagnosis, and personalizing treatment strategies. AI algorithms can efficiently analyze vast amounts of patient data from diverse sources, enabling accurate and timely diagnoses. When combined with CRISPR gene-editing technology, AI can identify the specific genetic changes responsible for COVID-19, leading to targeted treatments tailored to individual patients. This approach has the potential to overcome

the challenges of disease heterogeneity and increase treatment efficacy. Studying artificial intelligence in the diagnosis and treatment of COVID-19 using CRISPR technology aims to improve patient outcomes, improve the quality of life for COVID-19 patients, and contribute to advances in our understanding and management of this complex disease

CRISPR is a versatile defense mechanism that targets viral DNA in microbes by utilizing endonuclease. In 1987, it was first discovered in *Escherichia coli*. Experts believe that these arrangements are necessary for the adaptable, insensible framework that serves as the foundation for the educational and demonstration phases of CRISPR. The precise genetic modifications that are tailored to particular cell types can be produced by modifying this framework.(Pan & Kraschel, 2018). CRISPR and CRISPR-associated proteins guard bacteria's and archaea's immune systems against intrusive DNA elements. CRISPR-Cas systems are divided into two groups, six types, and 21 subtypes.(Hidalgo-Cantabrana et al., 2019). For use in animal and cell models, numerous targeted gene editing methods, such as ZFNs and TALENs, have been developed. With CRISPR/Cas, gene editing can be done quickly, easily, and effectively. (Mehravari et al., 2019). Numerous species' genomes can be altered easily and effectively using this strategy. (Zhan et al., 2019).

2.2 Detailed explanation of the CRISPR-Cas system.

In three phases, the CRISPR-Cas system defends against viruses and foreign genetic material. After the Cas proteins are produced, a spacer is transcribed into pre-crRNA. Accordingly, pre-crRNA is severed by Cas proteins into mature crRNA. In the wake of being integrated into the host's CRISPR locus, the protospacers act as stops between rehashes of the crRNA. Thirdly, the Cas protein starts genomic breaks after crRNA focusing on. Sequence-specific PAMs near specific crRNA sites in the target genome are required for multiple CRISPR systems to function.(Bengio et al., 2021).

The CRISPR-Cas system, an essential tool in genome editing technologies and medicinal research, is derived from the adaptive immune systems of bacteria and archaea.(Wright et al., 2016).

The CRISPR/Cas system uses three steps to get ready for infections and foreign genetic material. After the production of Cas proteins, a spacer is translated into pre-cr-RNA. Cas proteins then cleave the pre-crRNA to convert it into mature crRNA. Subsequently, the

protospacers are incorporated into the host's CRISPR locus to act as gaps between repeats of the crRNA. Thirdly, the Cas protein causes genomic breaks after cr RNA targeting. The presence of sequence-specific PAMs adjacent to particular crRNA sites in the target genome is necessary for a number of CRISPR systems.(Bengio et al., 2021).

In genome editing, double-stranded DNA at a particular region on the genome is broken using CRISPR-Cas9. The simplest approach is type II targeted nuclease, which uses CRISPR RNA and trans-activating CRISPR RNA to transfer the Cas9 nuclease to the target site. There is no impact from the fictitious merging of crRNA and tracrRNA into a single RNA chain sgRNA.(Ishino et al., 2018).

2.3 Discussion on its various applications beyond genome editing.

Recent research has demonstrated that when the ssDNA sequence that is not complementary to crRNA is present, extra cleavage phenomena is generated, leading to fast and complete cleavage of the ssDNA strand (Bonini et al., 2021; Chen et al., 2018; Gootenberg et al., 2018). Moreover, the cleavage is unrelated to the particular sequence of dsDNA. It is technically true that self-cleavage activation does not require an extra PAM sequence. The RuvC domain becomes accessible following target dsDNA unwinding and cleavage during the typical CRISPR-Cas12a action, while the non-target ssDNA cleavage happens on its own. using PCR and a fluorescent-quencher ssDNA reporter to examine the Cas12a/crRNA complex and its collateral activity, and creating a one-hour Low-cost Multipurpose highly Effector System (HOLMES). When used in genotype detection tests on human 293 T cells, Holmes was able to differentiate between homozygous and heterozygous genotypic variants (Li, Cheng, et al., 2018). A CRISPR-Cas12 assay for SARS-CoV-2 detection was developed, generating target amplicons through RT-RPA and identifying them through a gRNA complex connected to the collateral cleavage activity of fluorophore-tagged probes, allowing detection via fluorescent measurement or visual method (Talwar et al., 2021). Similarly, a CRISPR-Cas12-based lateral flow assay for the quick (less than 40 minutes), simple to use, and reliable detection of SARS-CoV-2 from respiratory swab RNA extracts has been reported. For RNA extracted from nasopharyngeal or oropharyngeal swabs in universal transport medium (UTM), this assay performs simultaneous reverse transcription and isothermal amplification using loop-mediated amplification (RT-LAMP). Cas12 detection of predefined coronavirus sequences is then performed, and cleavage of a reporter molecule confirms detection of the virus(Broughton et al., 2019). Therefore, this trans-ssDNA-cleavage activity of CRISPR-Cas12a provides a novel

approach to enhance transcription and replication responses in vivo, create more rapid, sensitive, and targeted tools for the identification of certain nucleic acid sequences.

2.4 Previous studies on CRISPR-Cas in genome engineering

Describes how a technique in molecular biology called CRISPR-Cas9 editing can change how a gene affects an organism's phenotype. Thanks to advances in genome editing and high-quality sequencing, the next generation of subatomic scientists focusing on science in today's homerooms should be able to comprehend the inherited variation of a diverse range of living organisms.(Thurtle-Schmidt & Lo, 2018)

2.5 Review of literature related to COVID-19 and genome editing

The SARS-CoV-2 epidemic has affected public health resources, endangered human health, and upended international economics. Genome editing became faster and less expensive with the discovery of the CRISPR/Cas system in 2012, a novel technique for changing the genomes of plants and animals. More and more, CRISPR/Cas is being used for disease detection and treatment because of its speed, cost, and precision. It can be helpful for researching coronavirus replication in cell culture and creating treatment approaches.(Hillary & Ceasar, 2023).

2.6 Examination of existing literature on CRISPR-Cas applications in combating COVID-19.

Machine learning-based genetic engineering applications are essential in the fight against COVID-19. In a recent experiment, Malone and his colleagues used the application. "The most successful therapeutic targets of immunotherapy are those that are visible on the cell surface and that T cells are most likely to recognize," with the goal of "assess which antigens possess the essential characteristics for HLA binding and modification." They also described the tools used to forecast immunogenicity and the way antigens are presented to infected patients, as well as the "whole SARS-CoV2 proteome" and specific hotspots inside the host cell. These results facilitate the prediction of broadly applicable techniques for pathogen-specific vaccination to the People of the World(Habehh & Gohel, 2021).

2.7 Deep Learning in Genomics

Artificial Intelligence (AI) is a branch of artificial intelligence that allows computers to reason without being specifically changed. Artificial Intelligence (AI) is used for a variety of processing tasks, and its main goal is to get the machine ready to solve a problem better by using the available data, which may be named in regulated learning and unlabeled in unassisted learning. The main goal is to teach computers to learn from their experiences (Das et al., 2015).

Deep learning algorithms often use the one-hot encoding of the DNA center succession to identify the properties of the target grouping as a result. Deep learning and machine learning algorithms can correctly forecast if gRNA will work against the target, but they can't capitalize on the potential that arises from integrating deep learning with physical chemistry and sequential characteristics. Proposing effective ways to estimate target performance with global information-gathering capacities remains difficult. (Xie et al., 2023).

2.7.1 Introduction to deep learning techniques and their applications in genomics.

These strict rules ensure that moral research is conducted properly and advances. One of the most challenging parts of employing machine learning (ML) in healthcare is interpreting and applying clinical data. The intricate architecture of machine learning (ML) techniques, especially deep learning-based ones, makes it very difficult to pinpoint and measure the original characteristics' contribution to the prediction. The lack of transparency has severely hindered the adoption of ML-based methods in the medical services sector; nevertheless, this is less of a problem in other ML applications (such as online search). The healthcare sector is fully aware of the importance of having direct access to the solution in addition to the solution itself. (Habehh & Gohel, 2021).

The DeepGuide deep learning framework leverages convolutional neural networks (CNNs) to enhance the performance of existing sgRNA activity prediction tools. To determine how the sgRNA scene is handled about the genome, a convolutional autoencoder was used for solo learning during the pre-preparing phase. Subsequently, the CNN was trained with controlled

learning using the grouping esteem and correlated chromatin availability data for every single sgRNA target site found in the Cas9 datasets. Cross-validating the model expectations allowed for the examination of the link between the actual and expected CS values. To verify the effectiveness of the recommended guidelines, an independent validation focused on a group of genes whose null mutations produce symptoms that are easily screenable. Using the Othdataset, DeepGuide accurately predicted 20 nt Cas9 sgRNA and outperformed previous guide movement expectation techniques. (Baisya et al., 2022).

For the identifiable evidence and measurement of the fundamental information components used in determining, a precise approach is anticipated. To increase adoption rates, machine learning techniques can also be created, put into practice, and assessed with the assistance of medical experts. Moreover, despite some concerns about the potential for less relational commitment between patients and PCPs as a result of the growing use of ML-based procedures, these tactics offer a remarkable possibility to further develop joining. Studies show that the idea of the doctor-patient connection is quickly disappearing, and about 25% of Americans do not have a primary care physician. (Habeheh & Gohel, 2021).

Profound CRISPR and Seq-deepCpf1 ~~25~~ are two examples of the sgRNA movement forecast devices that the DeepGuide convolutional neural network assembled profound learning structures. Using a convolutional autoencoder, unsupervised learning was performed in the pre-training phase to determine how the genome's sgRNA landscape is represented. Next, administered learning on the CNN was carried out using the configuration, CS esteem, and associated chromatin openness data for every sgRNA target site from the Cas9 datasets. Finally, by cross-validating the model predictions, the correlation between the actual and expected CS values was discovered. The efficacy of the suggested guidelines was independently confirmed by concentrating on a set of genes whose null mutations result in symptoms that are easy to screen for. In terms of predicting guide activity on the dataset, DeepGuide fared better than earlier approaches, correctly predicting 20 nt Cas9 sgRNA with NGG PAM and 25 nt Cas sgRNA with TTTV PAM (Baisya et al., 2022).

Complex off-target prediction models have been studied in the past, however they do not make good use of sequence pair information. The problem of really using arrangement pair data continue to be problematic. Forecasting is a problem that can be divided into two challenges related to deep learning. Make a vector or raster graphic that shows the sequence pair for gRNA

and DNA. To advance high-request highlights from vector or grid structure and generate expectations for grouping links, apply a deep learning model. (Z. R. Zhang & Jiang, 2022).

2.7.2 Review of studies using deep learning for analyzing genetic data.

Deep learning algorithms often use one-hot encoding of the DNA core sequence to automatically identify the properties of the target sequence. Even if deep learning and AI algorithms can accurately determine the viability of gRNA against the objective, they do not, in any event, take advantage of the potential that arises from coordinating deep learning with subsequent qualities and real science. Proposing effective techniques to anticipate goal execution with global data collection capabilities is still a work in progress. (Xie et al., 2023).

Ongoing advancements in profound learning-based protein structure expectation tactics, which make use of the complexity of these macromolecules to stifle, guide, or even modify certain illness-causing proteins, give new avenues for profoundly customized programs. Small proteins called anti-CRISPR proteins, derived from bacteriophages, provide defense against the CRISPR-Cas system's prokaryotic adaptive immunity. Here, using precise building expectations and real-world experiments, we demonstrate the various interference strategies exhibited by these (anti)CRISPR proteins.(Park et al., 2022).

Sequence programming is necessary for genome editing techniques. Inner fix processes are strengthened when site-explicit single-strand breaks or double-strand breaks are produced by site-explicit endonucleases at the specified location. fill in the gaps. (Rudin et al., 1989).

AI is being used with CRISPR to an ever-greater extent, and new assumption devices are always emerging. While the great majority of models support the conclusions of CRISPR-Cas9 research, there are some notable differences between them. While some models are straightforward and applicable to all organic entities and cell types, others are more intricate, making use of data such as epigenetic information to predict differences in CRISPR reasonability under different circumstances. Reasonable AI reduces the likelihood of human tendency by enhancing natural framework recognition with additional data and a wider range of environments. (O'brien et al., 2021).

Dynamic learning is a semi-supervised approach that labels unlabeled data by using a learning formula. A predetermined set of named data is used to set up an operational learning framework. The computation for the unlabeled data then predicts the most pertinent names. This method is especially intriguing in the scientific domain because collecting precise data can occasionally be expensive and time-consuming. While there is an abundance of highly modified unlabeled data and manual tagging is impractical, dynamic learning can be used to modify genomes. (Sherkatghanad et al., 2023).

Exploratory proficiency results generally don't match expectations very well, and the relapse models used to evaluate CRISPR productivity aren't very precise. The goal is to predict efficacy. due to the possibility that models with a regrettable awareness of assumptions could arise from the complex process of representing organic structures. Creating an incredibly persistent volume of sgRNA is predicted to require more data than just a throughput categorization. Small sample sizes and even fragmented imputation sets imply that fixed predictions would yield less significant findings than high/low classifications. Therefore, even if it is only a temporary solution until relapse calculations can demonstrate sgRNA efficiency, the limit of order calculations to distinguish between destinations that are more and less dynamic is helpful in practical applications. (O'Brien et al., 2021).

The benefits of CNN and LSTM are their ability to learn rational facts despite the tendencies of neighboring examples. We shall start with the optimal configuration of physicochemical features and work our way up to the brain network that drives the critical succession. We fully utilized the multimodal information from both branches by employing a direct element combination technique. Regression predictions that are executed on target allow for the binary classification of active and inactive gRNAs based on these predictions. CRISPR may also be used efficiently with a range of CRISPR systems and animals through transfer learning. (Xie et al., 2023).

The neighborhood-establishing data branch uses the equal design of a bidirectional long momentary memory organization and a convolutional brain organization to separate sophisticated grouping highlights from the basic succession DNA. The exactness of expectancies can be greatly increased by selecting a determination of physicochemical elements with RNA optional construction strengthened to add additional data to the neural network. Given the combination of multi-layered sequencing properties and visual and auditory highlights, more precise data can be recovered. The CRISPR-Cas9 framework can be improved using a technique called transfer learning. CRISPR outperforms other developments in creature

models and diverse CRISPR frameworks. Consequently, it offers a simple framework for highly accurate and broadly applicable targeting efficiency prediction.

A third-generation gene editing technique that is widely used in biological applications is the CRISPR/Cas9 system. One difficult problem facing CRISPR/Cas9 technology in real-world applications is (off-target) consequences. Although many prediction models have been developed to predict off-target behaviors, the existing methods do not make proper use of sequence pair data. There is still room for more accuracy. The CRISPR-IP model learns sequence pair features via CNN, BiLSTM, and the attention layer. Performance evaluations on two datasets show that our encoding method can represent sequence pair information accurately and that the CRISPR-IP model performs better than other models. These techniques can recognize sequence pair traits automatically and predict results based on those attributes. (Z. R. Zhang & Jiang, 2022) .

2.8.2 Review of studies exploring indel frequencies in various contexts.

Many design methods have been developed, but the ability to reliably forecast between different species and Cas enzyme types is lacking. The use of the CRISPR framework is largely dependent on the viability of sgRNA. Most prediction algorithms are trained using data from a small number of species, most commonly E. Coli or human and mouse cell lines. This poses a crucial challenge. Furthermore, Cas9 variants have been used in the majority of screens that have linked sgRNA groups to activity thus far. Homing activity for other species cannot be predicted based on the strong association between the sgRNA features and the target species' homing activity; instead, the changeability between species may be caused by variations in the genomic scene. (Moreb & Lynch, 2021)

CHAPTER III

Methodology

PREDICTIVE DATA MODELING OF CRISPR-CAS12 USING VARIOUS ENCODING METHODOLOGIES

3.1 INTRODUCTION

This work is carefully planned to analyze and comprehend how three different DNA sequence encoding methods—K-mers, One-hot, and Integer Encoding—affect the prediction abilities of four different machine learning models: Support Vector Regression(SVR), Decision Tree, Random Forest, and XGBoost. Estimating the frequency of insertions and deletions (indels), which are essential elements of genome engineering regulated by the CRISPR-Cas12 (Cpf1) system, is the main focus. Because indels are essential to genome editing and provide information about the function and regulation of genes, it is critical to accurately predict indels for genetic research and therapeutic interventions. The goal of this thorough analysis is to provide insight into the best methods for encoding genomic sequences and the best machine learning models to use in order to forecast CRISPR-Cas12 induced indel frequencies. The results are anticipated to have wider ramifications in the fields of genetics, bioinformatics, and medicinal development in addition to offering useful suggestions for practitioners and researchers in the field of genome editing.

Support Vector Regressor

Support Vector Regressor (SVR) is regarded as one of the most popular and efficient classifiers. The principal mechanism behind the classification approach of SVRR revolves around the construction of hyperplane or set of hyperplanes in high dimension space. SVR can be classified into linear and non-linear. Linear SVR algorithms is develop using a technique known as kernel trick which is a function that transform high dimensional space from low dimensional space inputs [32]. Given a set of training data $\{(x_i, d_i)\}_i^N$ (d_i is the actual value, x_i represents the input vector and N is the data number), given that the SVM function is:

$$y = f(x) = w\phi(x_i) + b \quad (1)$$

When input vector , x , which are input feature spaces, is non-linearly mapped to $\phi(X)$.

Then, the SVR equation is given as

$$f(x, \alpha_j, \alpha_{i^*}) = \sum_i^N (\alpha_i - \alpha_{i^*}) K(x, x_i) + b \quad (2)$$

$k(x_i, x_j)$ is the bias term, and is the kernel function in the feature space following non-linear mapping. The most widely used kernel function is the Gaussian Radial Basis Function (RBF), which is easier to use and performs better than both linear and polynomial kernels when mapping non-linear training data into infinite-dimensional space. It is expressed as follows:

$$k(x_1, x_2) = \exp(-\gamma \|x_1 - x_2\|^2) \quad (3)$$

where γ is the kernel parameter.

Decision Trees

Decision Trees (DT) break down complex solutions into smaller, more manageable options in order to organize them in a structure like a tree. Decision trees employ splits to choose attributes that reduce entropy in order to provide accurate class assignments. The model's visual depiction and feature importance insights make it easier to understand. Pruning techniques lessen overfitting while ensembles like Random Forests and Boosting increase prediction accuracy. (Mienye et al., 2019).

$$f(x) = \begin{cases} c_1 & \text{if } x_i \leq T_i \\ c_2 & \text{if } x_i > T_i \end{cases}$$

$F(x)$ is the prediction made by the decision tree.

X_i is the i th feature.

T_i is the threshold for the i -th feature.

c_1 and c_2 are the predicted classes or values for the corresponding branches.

Random Forest

The approach of group learning Random Forests is an effective tool. By using many decision trees and a random selection of attributes and data points, random forest (RF) minimizes overfitting and enhances generalization. The decision tree's collective intelligence is taken into account during prediction through voting or averages. This approach, which is well known for

its flexibility and interpretability and performs well with complex or noisy data, highlights the significance of the attributes.(Pavlov, 2019).

$$f(x) = \frac{1}{N} + \sum_{i=1}^N f_i(x)$$

$f(x)$ is the prediction made by the random forest.

N is the number of trees in the forest.

$f_i(x)$ is the prediction made by the i -th decision tree.

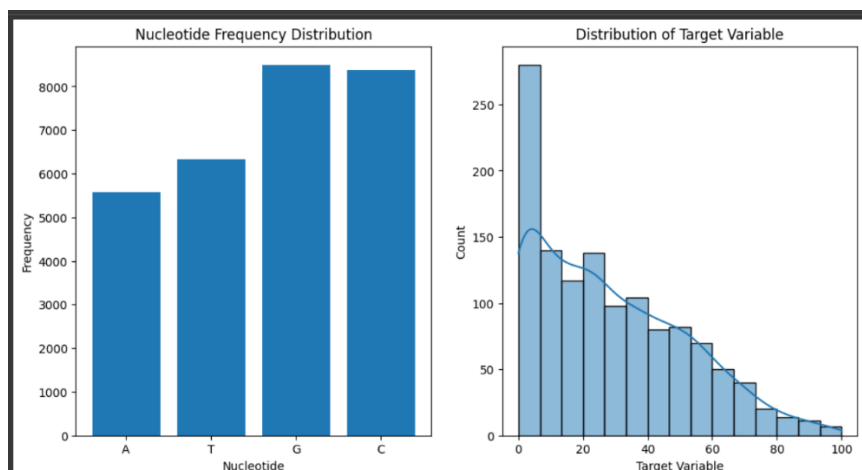
XGBoost

XGBoost is an optimized gradient-boosting algorithm that focuses on reducing errors iteratively. It builds decision trees sequentially, with each tree correcting the errors of the previous ones.

The study will employ the **Spearman Correlation coefficient** as the primary metric for evaluating the models' performance. This non-parametric measure will assess the rank-order relationship between the predicted and actual indel frequencies, providing insights into the models' ability to capture the underlying trends in the data, rather than just the absolute values.

3.2 THE DATASET DESCRIPTION

The primary dataset used in this study was provided by (Kim et al., 2017). Kim and associates searched for sgRNA characteristics associated with CRISPR-Cas12a activity.



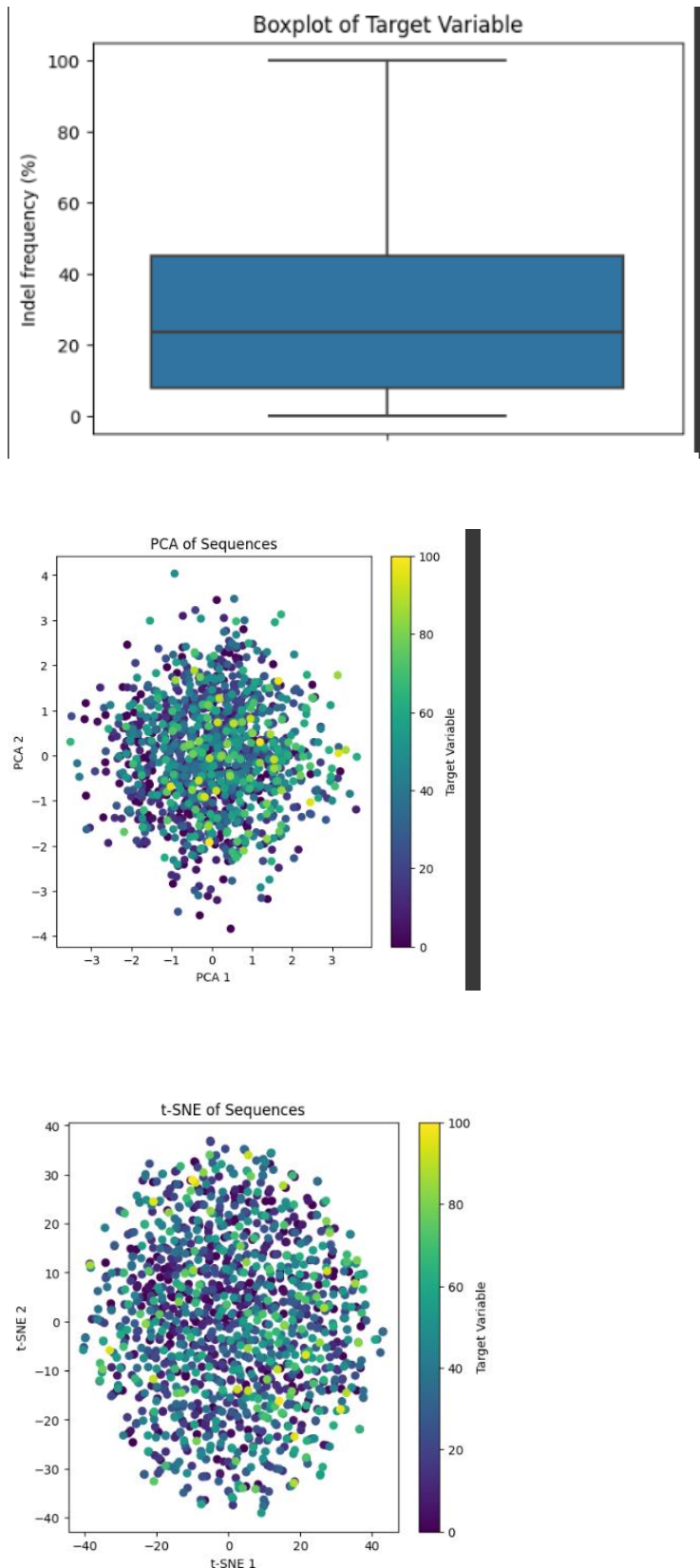


Figure 4.1 data set description.

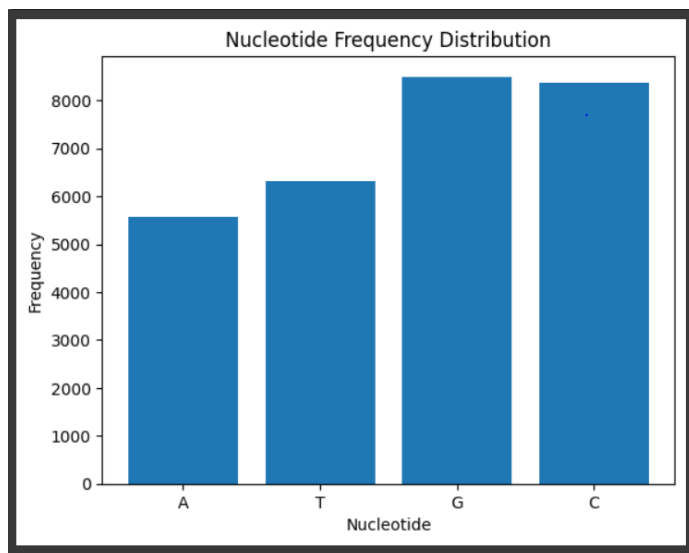


Figure 5.1 visualization of nucleotide frequency.

3.3 THE ENCODING TECHNIQUES

3.3.1 K-mers Encoding

Encoding Technique: K-mer encoding using CountVectorizer. K-mers are length-k subsequences that are taken out of DNA sequences. These k-mer sequences are transformed into feature vectors by the CountVectorizer, which counts the instances of each k-mer.

3.3.2 One-hot Encoding

Encoding technique: One-hot encoding using an established code dictionary is the encoding technique used. Each nucleotide is represented as a binary vector of length four, where each place signifies one of the nucleotides (A, T, G, and C). Every nucleotide is mapped to a matching binary vector by the coding dictionary. Then, a concatenation of these binary vectors represents the full sequence of DNA.

3.3.3 Integer Encoding

Encoding Technique: Integer encoding based on a predefined code dictionary. A specified code dictionary is used to represent each nucleotide (A, T, G, and C) by an integer number. This represents the whole sequence of DNA as a series of these integer numbers.

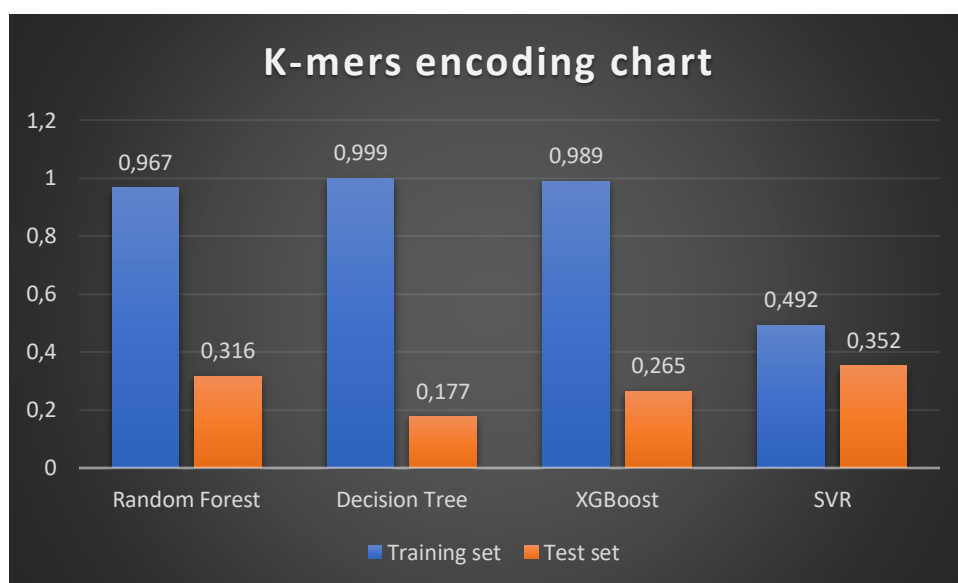
3.4 RESULTS AND DISCUSSION

Using the CRISPR-CAS12, I trained various regression models (Random Forest, Decision Tree, XGBoost, and SVR) on the encoded data, and then I used the Spearman correlation on training and test sets to assess each model's performance. Below we will see each one of the methods used above (k-mers encoding, one-hot encoding, and integer encoding) to train performance of the models:

1) K-mers encoding:

	Random Forest	Decision Tree	XGBoost	SVR
Training set	0.967	0.999	0.989	0.492
Test set	0.316	0.177	0.265	0.352

Table 3.1 : K-mers encoding.



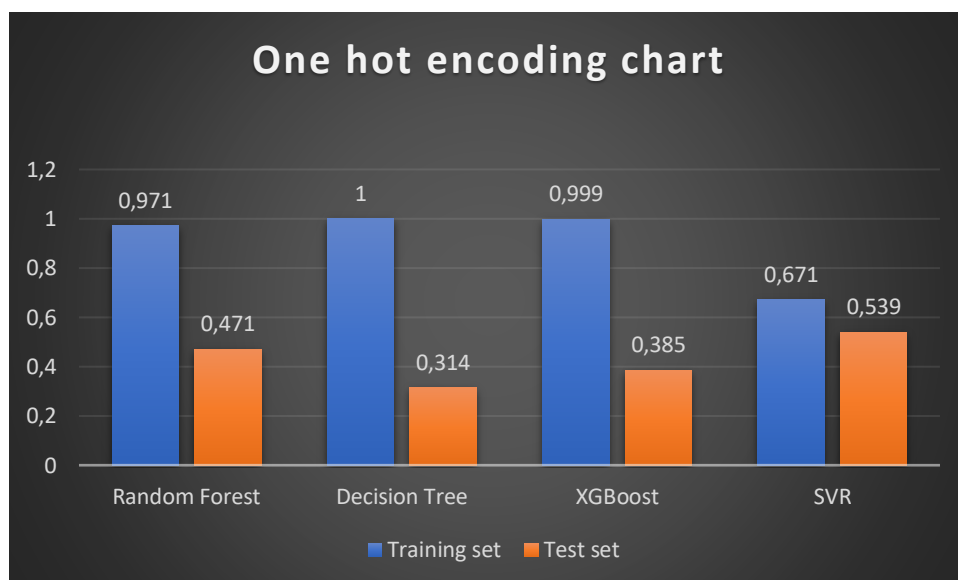
In this K-mers encoding technique we see that the training set of decision tree comes on top with a 0.999 spearman correlation which shows how well the decision tree model can identify

patterns in the training set of data but the test set is 0.177 spearman correlation which implies that when the model is used with fresh, untested data, its performance drastically drops.

2) One-hot encoding :

	Random Forest	Decision Tree	XGBoost	SVR
Training set	0.971	1.000	0.999	0.671
Test set	0.471	0.314	0.385	0.539

Table 3.2 : One-hot encoding.

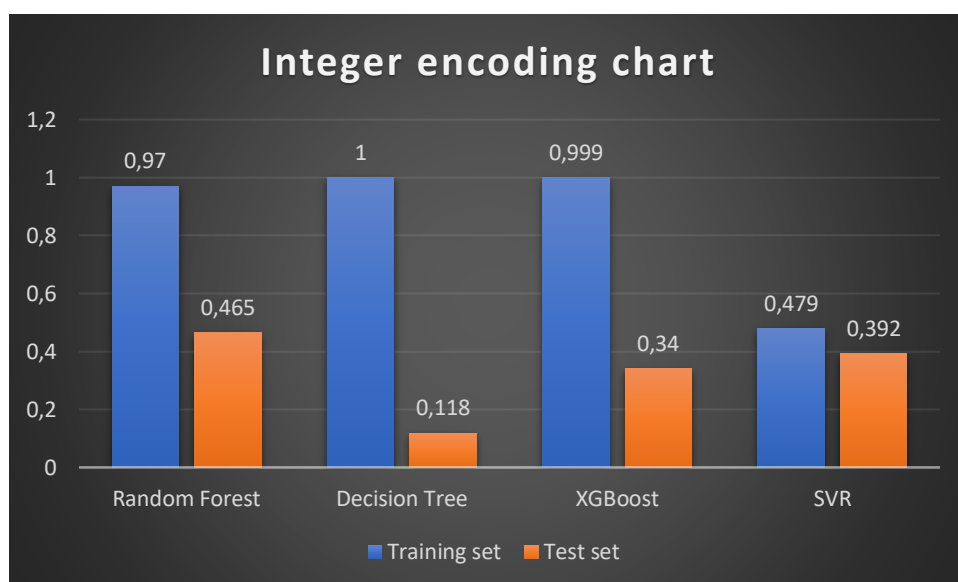


In this One-hot encoding technique we found out that the decision tree training set has the highest correlation of 1.000 spearman correlation which implies that the training set's data order is perfectly captured by the model but the test set has a spearman correlation of 0.314 which means it's not as strong as the perfect correlation of the training set

3) Integer encoding:

	Random Forest	Decision Tree	XGBoost	SVR
Training set	0.970	1.000	0.999	0.479
Test set	0.465	0.118	0.340	0.392

Table 3.3 : Integer encoding.



The decision tree training set in this Integer encoding technique has the highest correlation of 1.000 spearman correlation which means the same as the one hot encoding technique which it say that the training set data order is perfectly captured by the model (decision tree) while the test set has a spearman correlation of 0.118 which implies that it's not as strong as the perfect correlation in the training set same like the one-hot encoding technique

3.5 CONCLUSION

Model performance on training sets seems to be usually good, but depending on the encoding method, the models' performance differs on test sets. Based on the test sets for the majority of models, one-hot encoding appears to provide better Spearman correlations, indicating that it could be a more successful encoding method for this particular dataset.

CHAPTER IV

Methodology

CAS12 CRISPR DATA: PREDICTION WITH VARIOUS DNA SEQUENCE ENCODINGS AND DATA SCALING

4.1 INTRODUCTION.

This work is meticulously planned to dissect and understand the impact of three distinct DNA sequence encoding techniques—K-mers, One-hot, and Integer Encoding—on the predictive capabilities of four sophisticated machine learning models: Support Vector Regression (SVR), Decision Tree, Random Forest, and XGBoost. The primary aim is to estimate the frequency of insertions and deletions (indels), crucial elements of genome engineering regulated by the CRISPR-Cas12 (Cpf1) system. Indels are central to genome editing, providing insights into gene function and regulation; thus, their precise prediction is crucial for advancing genetic research and therapeutic interventions.

After the encoding process, **Data Scaling** will be applied to standardize the range of the feature data. This step is crucial as it ensures that the numerical values of the features have been normalized, allowing the machine learning models to converge more quickly during training and reducing the chance of bias towards certain features due to their scale. Various scaling techniques, such as Min-Max scaling, Standard scaling, or Robust scaling, might be employed depending on the distribution and nature of the encoded data. The selection of an appropriate scaling method will be based on preliminary analysis and the specific characteristics of each encoding technique.

Support Vector Regressor

Support Vector Regressor (SVR) is regarded as one of the most popular and efficient classifiers. The principal mechanism behind the classification approach of SVRR revolves around the construction of hyperplane or set of hyperplanes in high dimension space. SVR can be classified into linear and non-linear. Linear SVR algorithms is develop using a technique known as kernel trick which is a function that transform high dimensional space from low dimensional space inputs [32]. Given a set of training data $\{(x_i, d_i)\}_i^N$ (d_i is the actual value, x_i represents the input vector and N is the data number), given that the SVM function is:

$$y = f(x) = w\phi(x_i) + b \quad (1)$$

When input vector , x , which are input feature spaces, is non-linearly mapped to $\phi(X)$.

Then, the SVR equation is given as

$$f(x, \alpha_j, \alpha_{i^*}) = \sum_i^N (\alpha_i - \alpha_{i^*})K(x, x_i) + b \quad (2)$$

$k(x_i, x_j)$ is the bias term, and is the kernel function in the feature space following non-linear mapping. The most widely used kernel function is the Gaussian Radial Basis Function (RBF), which is easier to use and performs better than both linear and polynomial kernels when mapping non-linear training data into infinite-dimensional space. It is expressed as follows:

$$k(x_1, x_2) = \exp(-\gamma\|x_1 - x_2\|^2) \quad (3)$$

where γ is the kernel parameter.

Decision Trees

Decision Trees (DT) break down complex solutions into smaller, more manageable options in order to organize them in a structure like a tree. Decision trees employ splits to choose attributes that reduce entropy in order to provide accurate class assignments. The model's visual depiction and feature importance insights make it easier to understand. Pruning techniques lessen overfitting while ensembles like Random Forests and Boosting increase prediction accuracy.(Mienye et al., 2019).

$$f(x) = \begin{cases} c_1 & \text{if } x_i \leq T_i \\ c_2 & \text{if } x_i > T_i \end{cases}$$

$F(x)$ is the prediction made by the decision tree.

X_i is the *ith* feature.

T_i is the threshold for the *i*-th feature.

c_1 and c_2 are the predicted classes or values for the corresponding branches.

Random Forest

The approach of group learning Random Forests is an effective tool. By using many decision trees and a random selection of attributes and data points, random forest (RF) minimizes overfitting and enhances generalization. The decision tree's collective intelligence is taken into

account during prediction through voting or averages. This approach, which is well known for its flexibility and interpretability and performs well with complex or noisy data, highlights the significance of the attributes.(Pavlov, 2019).

$$f(x) = \frac{1}{N} + \sum_{i=1}^N f_i(x)$$

$f(x)$ is the prediction made by the random forest.

N is the number of trees in the forest.

$f_i(x)$ is the prediction made by the i -th decision tree.

XGBoost

XGBoost is an optimized gradient-boosting algorithm that focuses on reducing errors iteratively. It builds decision trees sequentially, with each tree correcting the errors of the previous ones.

The study will employ the **Spearman Correlation coefficient** as the primary metric for evaluating the models' performance. This non-parametric measure will assess the rank-order relationship between the predicted and actual indel frequencies, providing insights into the models' ability to capture the underlying trends in the data, rather than just the absolute values.

4.2 THE DATASET DESCRIPTION

The primary dataset used in this study was provided by (Kim et al., 2017). Kim and associates searched for sgRNA characteristics associated with CRISPR-Cas12a activity.

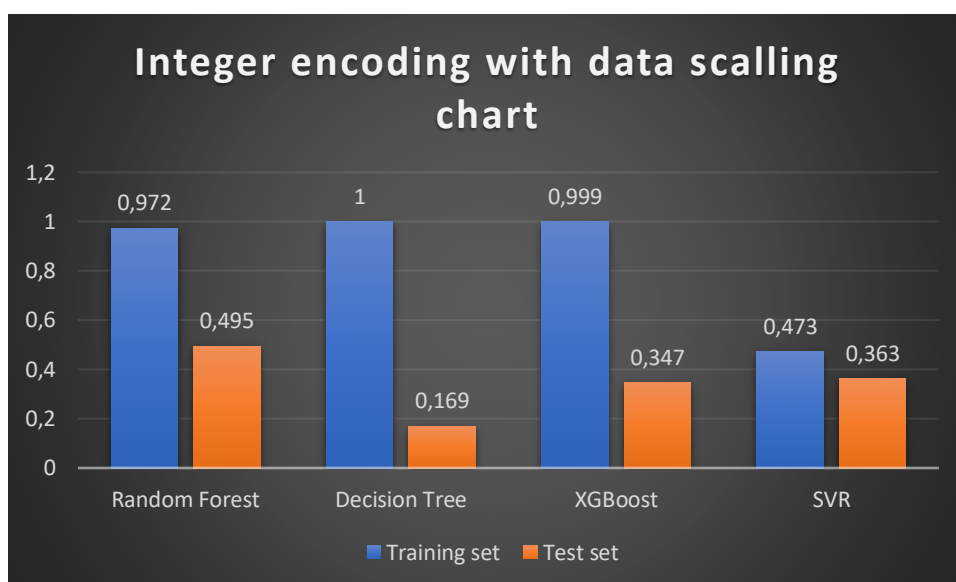
4.3 RESULTS AND DISCUSSION

I trained many regression models (Random Forest, Decision Tree, XGBoost, and SVR) using encoded DNA sequences using the CRISPR-CAS12 technology. Data scaling was used together with the encoding techniques, which included integer, one-hot, and k-mers. Through the use of Spearman correlation on training and test sets, performance evaluation provided useful information into each model's predictive capacity. The investigation explores how encoding schemes, data scaling, and regression models affect CRISPR-CAS12 applications, ranging from numeric representation in integer encoding to nuanced correlations in one-hot

encoding and unique k-mer characteristics. . Below we will see each one of the methods used above (integer encoding, one-hot encoding and k-mers encoding) to train performance of the models:

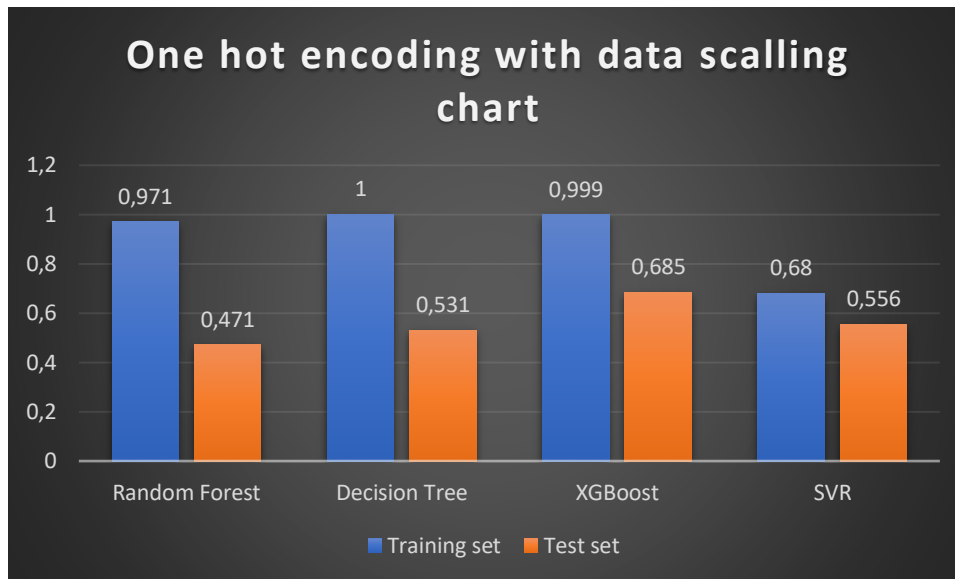
1) Integer Encoding with Data Scaling

Models	Random Forest	Decision Tree	XGBoost	SVR
Training set	0.972	1.000	0.999	0.473
Test set	0.495	0.169	0.347	0.363



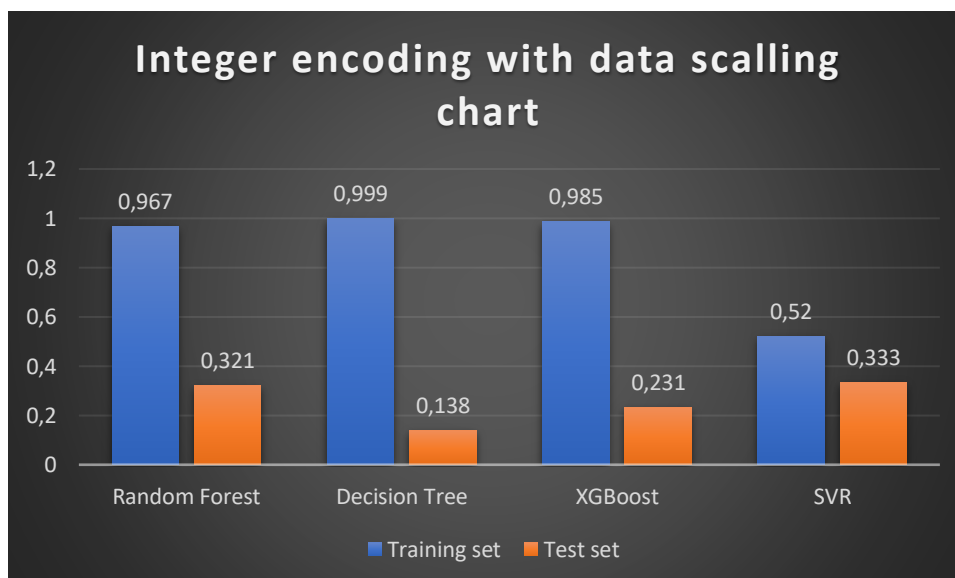
2) One-Hot Encoding with Data Scaling

Models	Random Forest	Decision Tree	XGBoost	SVR
Training set	0.971	1.00	0.999	0.680
Test set	0.471	0.531	0.685	0.556



3) Integer Encoding with Data Scaling

Models	Random Forest	Decision Tree	XGBoost	SVR
Training set	0.967	0.999	0.985	0.520
Test set	0.321	0.138	0.231	0.333



4.4 CONCLUSION

On the training set, Decision Tree consistently displays perfect correlation, which could be a sign of overfitting. Maintaining nucleotide relationships appears to be important since one-hot encoding appears to perform better than integer encoding. In general, RandomForest and XGBoost exhibit strong performance in various encoding methods. The efficacy of SVR varies depending on the encoding method selected. Performance on the test set is typically worse than on the training set, suggesting that certain models may be overfitting.

CHAPTER VI

Conclusion and Recommendations

Genomic research, agricultural practices, and medical fields have all made use of CRISPR, an enzyme and guide RNA system that recognizes and cuts a target DNA sequence. Selecting the right guide RNAs (gRNAs) is made easier by computational methods. Guide RNAs are necessary to initiate the CRISPR editing step. Various strategies have been used to develop CRISPR sgRNA design and evaluation tools, especially for the CRISPR/Cas9 system.

Gene editing is approached differently in CRISPR Cas12 systems. In contrast, Cas12 generates sharp-ended breaks upon detecting G-rich PAM. Cas12 needs a tight PAM sequence and employs sgRNA. The particular requirements of gene editing determine which Cas12 to use. Every system has benefits and is necessary for the advancement of CRISPR technology.

In conclusion, As seen in this first case, the performance of the various machine learning models on the training sets consistently yields positive outcomes. Nonetheless, the coding method selection had a major impact on how well the models performed on the test sets. Crucially, hot encoding was discovered to be a successful tactic on the test sets, exhibiting enhanced Spearman correlations and indicating that it functions effectively on this particular CRISPR dataset.

In the second case, the decision tree model's training dataset with the data measure regularly displayed perfect correlation, which increased the risk of overfitting. Nucleotide link preservation is crucial, as evidenced by hot encoding's superior performance over integer encoding. XGBoost and RandomForest frequently offer good performance across a range of encryption algorithms when it comes to managing nucleotide data. However, the efficiency with which support vector regression (SVR) may be carried out depends on the coding technique selected.

In the latter study, A consistent pattern of lower performance in the test set relative to the training set raised the possibility that some models were overfitting. This latter case is employed in several research projects, ranging from direct nucleotide frequency displays to the development and assessment of intricate machine learning models. A variety of feature representations, including hot encryption, k-mer counting, and integer encryption, are thoroughly examined. Low-dimensional datasets are also easier to visualize because to dimensionality reduction techniques, and the models were assessed using the Spearman correlation grading scale. Based on the results shown, it appears that the Random Forest model

is the most stable and dependable for this particular task and dataset.

Recommendations

There are still a few intriguing avenues to explore, even though the deep learning models created in this thesis have demonstrated their value in developing sgRNA for CRISPRcas12 and have enhanced the performance of sgRNA activity prediction. Experimenting with different deep learning-based architectures and employing better methods for optimum hyperparameter selection are two ways to potentially increase performance. Finding a way to increase the amount of features that can be utilized to build the model is the second challenge. This thesis only employed sgRNA sequence data; however, to improve model performance, epigenetic characteristics can be added. Moreover, integrating activity prediction and sgRNA off-target prediction into a single model would provide thorough information for choosing appropriate sgRNAs for CRISPR/Cas12 gene editing.

References

- Ameen, Z. S. id, Ozsoz, M., Mubarak, A. S., Turjman, F. Al, & Serte, S. (2021). C-SVR Crispr: Prediction of CRISPR/Cas12 guideRNA activity using deep learning models. *Alexandria Engineering Journal*, 60(4), 3501–3508. <https://doi.org/10.1016/j.aej.2021.02.007>
- Baisya, D., Ramesh, A., Schwartz, C., Lonardi, S., & Wheeldon, I. (2022). Genome-wide functional screens enable the prediction of high activity CRISPR-Cas9 and -Cas12a guides in *Yarrowia lipolytica*. *Nature Communications*, 13(1). <https://doi.org/10.1038/s41467-022-28540-0>
- Bengio, Y., Lecun, Y., & Hinton, G. (2021). Deep learning for AI. *Communications of the ACM*, 64(7), 58–65. <https://doi.org/10.1145/3448250>
- Bonini, A., Poma, N., Vivaldi, F., Kirchhain, A., Salvo, P., Bottai, D., Tavanti, A., & Di, F. (2021). Advances in biosensing : The CRISPR / Cas system as a new powerful tool for the detection of nucleic acids. *Journal of Pharmaceutical and Biomedical Analysis*, 192, 113645. <https://doi.org/10.1016/j.jpba.2020.113645>
- Broughton, J. P., Deng, X., Yu, G., Fasching, C. L., Servellita, V., Singh, J., Miao, X., Streithorst, J. A., Granados, A., Sotomayor-gonzalez, A., Zorn, K., Gopez, A., Hsu, E., Gu, W., Miller, S., Pan, C., Guevara, H., Wadford, D. A., Chen, J. S., & Chiu, C. Y. (2019). CRISPR – Cas12-based detection of SARS-CoV-2. *Nature Biotechnology*, 38, 870–874. <https://doi.org/https://doi.org/10.1038/s41587-020-0513-4>
- Cascella, M., Rajnik, M., & Aleem, A. (2022). Features, Evaluation, and Treatment of Coronavirus (COVID-19). *National Library of Medicine*, 2019(November), 30. <https://www.ncbi.nlm.nih.gov/books/NBK554776/>
- Chen, J. S., Chen, J. S., Ma, E., Harrington, L. B., Costa, M. Da, Tian, X., & Palefsky, J. M. (2018). *CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity*. 6245(February), 1–8.
- Das, S., Dey, A., Pal, A., & Roy, N. (2015). Applications of Artificial Intelligence in Machine Learning: Review and Prospect. *International Journal of Computer Applications*, 115(9), 31–41. <https://doi.org/10.5120/20182-2402>
- Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24–29. <https://doi.org/10.1038/s41591-018-0316-z>
- Fonfara, I., Richter, H., BratoviÄ, M., Le Rhun, A., & Charpentier, E. (2016). The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature*, 532(7600), 517–521. <https://doi.org/10.1038/nature17945>
- Friedrichs, S., Takasu, Y., Kearns, P., Dagallier, B., Oshima, R., Schofield, J., & Moreddu, C. (2019). An overview of regulatory approaches to genome editing in agriculture. *Biotechnology Research and Innovation*, 3(2), 208–220. <https://doi.org/10.1016/j.biori.2019.07.001>
- Gootenberg, J. S., Abudayyeh, O. O., Kellner, M. J., Joung, J., Collins, J. J., & Zhang, F. (2018). Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a and Csm6. *Science*, 360(6387), 439–444. <https://doi.org/10.1126/science.aag0179>
- Gostimskaya, I. (2022). CRISPR–Cas9: A History of Its Discovery and Ethical Considerations of Its Use in Genome Editing. *Biochemistry (Moscow)*, 87(8), 777–788. <https://doi.org/10.1134/S0006297922080090>
- Habebh, H., & Gohel, S. (2021). Machine Learning in Healthcare. *Current Genomics*, 22(4), 291–300. <https://doi.org/10.2174/1389202922666210705124359>
- Hidalgo-Cantabrana, C., Goh, Y. J., & Barrangou, R. (2019). Characterization and

- Repurposing of Type I and Type II CRISPR–Cas Systems in Bacteria. *Journal of Molecular Biology*, 431(1), 21–33. <https://doi.org/10.1016/j.jmb.2018.09.013>
- Hillary, V. E., & Ceasar, S. A. (2023). A Review on the Mechanism and Applications of CRISPR/Cas9/Cas12/Cas13/Cas14 Proteins Utilized for Genome Engineering. *Molecular Biotechnology*, 65(3), 311–325. <https://doi.org/10.1007/s12033-022-00567-0>
- Hille, F., Richter, H., Wong, S. P., Bratovič, M., Ressel, S., & Charpentier, E. (2018). The Biology of CRISPR-Cas: Backward and Forward. *Cell*, 172(6), 1239–1259. <https://doi.org/10.1016/j.cell.2017.11.032>
- Ishino, Y., Krupovic, M., & Forterre, P. (2018). History of CRISPR-Cas from Encounter with a Mysterious. *Journal of Bacteriology*, 200(7), e00580-17.
- Jackson, S. A., McKenzie, R. E., Fagerlund, R. D., Kieper, S. N., Fineran, P. C., & Brouns, S. J. J. (2017). CRISPR-Cas: Adapting to change. *Science*, 356(6333). <https://doi.org/10.1126/science.aal5056>
- Kass, E. M., & Jasin, M. (2014). *Break Repair Pathways*. 584(17), 3703–3708. <https://doi.org/10.1016/j.febslet.2010.07.057>. Collaboration
- Kim, H. K., Song, M., Lee, J., Menon, A. V., Jung, S., Kang, Y. M., Choi, J. W., Woo, E., Koh, H. C., Nam, J. W., & Kim, H. (2017). In vivo high-throughput profiling of CRISPR-Cpf1 activity. *Nature Methods*, 14(2), 153–159. <https://doi.org/10.1038/nmeth.4104>
- Kim, H., Kim, S. T., Ryu, J., Kang, B. C., Kim, J. S., & Kim, S. G. (2017). CRISPR/Cpf1-mediated DNA-free plant genome editing. *Nature Communications*, 8, 1–7. <https://doi.org/10.1038/ncomms14406>
- Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
- Li, S., Cheng, Q., Wang, J., Zhao, G., & Wang, J. (2018). CRISPR-Cas12a-assisted nucleic acid detection. *Cell Discovery*, 18–21. <https://doi.org/10.1038/s41421-018-0028-z>
- Li, S., Li, J., Zhang, J., Du, W., Fu, J., Sutar, S., Zhao, Y., & Xia, L. (2018). Synthesis-dependent repair of Cpf1-induced double strand DNA breaks enables targeted gene replacement in rice. *Journal of Experimental Botany*, 69(20), 4715–4721. <https://doi.org/10.1093/jxb/ery245>
- Matsumoto, D., & Nomura, W. (2023). The history of genome editing: advances from the interface of chemistry & biology. *Chemical Communications*, 59(50), 7676–7684. <https://doi.org/10.1039/d3cc00559c>
- Mehravar, M., Shirazi, A., Nazari, M., & Banan, M. (2019). Mosaicism in CRISPR/Cas9-mediated genome editing. *Developmental Biology*, 445(2), 156–162. <https://doi.org/10.1016/j.ydbio.2018.10.008>
- Mienye, I. D., Sun, Y., & Wang, Z. (2019). Prediction performance of improved decision tree-based algorithms: a review. *Procedia Manufacturing*, 35, 698–703. <https://doi.org/10.1016/j.promfg.2019.06.011>
- Moreb, E. A., & Lynch, M. D. (2021). Genome dependent Cas9/gRNA search time underlies sequence dependent gRNA activity. *Nature Communications*, 12(1). <https://doi.org/10.1038/s41467-021-25339-3>
- Moreno-Mateos, M. A., Vejnár, C. E., Beaudoin, J. D., Fernandez, J. P., Mis, E. K., Khokha, M. K., & Giraldez, A. J. (2015). CRISPRscan: Designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nature Methods*, 12(10), 982–988. <https://doi.org/10.1038/nmeth.3543>
- O'Brien, A. R., Burgio, G., & Bauer, D. C. (2021). Domain-specific introduction to machine learning terminology, pitfalls and opportunities in CRISPR-based gene editing. *Briefings in Bioinformatics*, 22(1), 308–314. <https://doi.org/10.1093/bib/bbz145>
- Pan, A., & Kraschel, K. L. (2018). CRISPR diagnostics: Underappreciated uses in

- perinatology. *Seminars in Perinatology*, 42(8), 525–530.
<https://doi.org/10.1053/j.semperi.2018.09.016>
- Park, H. M., Park, Y., Vankerschaver, J., Van Messem, A., De Neve, W., & Shim, H. (2022). Rethinking Protein Drug Design with Highly Accurate Structure Prediction of Anti-CRISPR Proteins. *Pharmaceuticals*, 15(3), 1–14. <https://doi.org/10.3390/ph15030310>
- Pavlov, Y. L. (2019). Random forests. *Random Forests*, 1–122.
<https://doi.org/10.1201/9780429469275-8>
- Rudin, N., Sugarman, E., & Haber, J. E. (1989). Genetic and physical analysis of double-strand break repair and recombination in *Saccharomyces cerevisiae*. *Genetics*, 122(3), 519–534. <https://doi.org/10.1093/genetics/122.3.519>
- Schmidt, S. M., Belisle, M., & Frommer, W. B. (2020). The evolving landscape around genome editing in agriculture. *EMBO Reports*, 21(6), 19–22.
<https://doi.org/10.15252/embr.202050680>
- Sherkatghanad, Z., Abdar, M., Charlier, J., & Makarenkov, V. (2023). Using traditional machine learning and deep learning methods for on- and off-target prediction in CRISPR/Cas9: a review. *Briefings in Bioinformatics*, 24(3), 1–25.
<https://doi.org/10.1093/bib/bbad131>
- Talwar, C. S., Park, K., Ahn, W., Kim, Y., Kwon, O. S., Yong, D., Kang, T., & Woo, E. (2021). Detection of Infectious Viruses Using CRISPR-Cas12-Based Assay. *Biosensors*, 11. <https://doi.org/10.3390/bios11090301>
- Thurtle-Schmidt, D. M., & Lo, T. W. (2018). Molecular biology at the cutting edge: A review on CRISPR/CAS9 gene editing for undergraduates. *Biochemistry and Molecular Biology Education*, 46(2), 195–205. <https://doi.org/10.1002/bmb.21108>
- Van Der Oost, J., Westra, E. R., Jackson, R. N., & Wiedenheft, B. (2014). Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nature Reviews Microbiology*, 12(7), 479–492. <https://doi.org/10.1038/nrmicro3279>
- Wright, A. V., Nuñez, J. K., & Doudna, J. A. (2016). Biology and Applications of CRISPR Systems: Harnessing Nature’s Toolbox for Genome Engineering. *Cell*, 164(1–2), 29–44. <https://doi.org/10.1016/j.cell.2015.12.035>
- Xie, J., Liu, M., & Zhou, L. (2023). CRISPR-OTE: Prediction of CRISPR On-Target Efficiency Based on Multi-Dimensional Feature Fusion. *Irbm*, 44(1), 100732.
<https://doi.org/10.1016/j.irbm.2022.07.003>
- Xu, H., Xiao, T., Chen, C. H., Li, W., Meyer, C. A., Wu, Q., Wu, D., Cong, L., Zhang, F., Liu, J. S., Brown, M., & Liu, X. S. (2015). Sequence determinants of improved CRISPR sgRNA design. *Genome Research*, 25(8), 1147–1157.
<https://doi.org/10.1101/gr.191452.115>
- Yan, J., Chuai, G., Zhou, C., Zhu, C., Yang, J., Zhang, C., Gu, F., Xu, H., Wei, J., & Liu, Q. (2018). Benchmarking CRISPR on-target sgRNA design. *Briefings in Bioinformatics*, 19(4), 721–724. <https://doi.org/10.1093/bib/bbx001>
- Yoshida, M., Saito, T., Takayanagi, Y., Totsuka, Y., & Onaka, T. (2022). Necessity of integrated genomic analysis to establish a designed knock-in mouse from CRISPR-Cas9-induced mutants. *Scientific Reports*, 12(1), 1–15. <https://doi.org/10.1038/s41598-022-24810-5>
- Zetsche, B., Gootenberg, J. S., Abudayyeh, O. O., Slaymaker, I. M., Makarova, K. S., Essletzbichler, P., Volz, S. E., Joung, J., Van Der Oost, J., Regev, A., Koonin, E. V., & Zhang, F. (2015). Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell*, 163(3), 759–771. <https://doi.org/10.1016/j.cell.2015.09.038>
- Zhan, T., Rindtorff, N., Betge, J., Ebert, M. P., & Boutros, M. (2019). CRISPR/Cas9 for cancer research and therapy. *Seminars in Cancer Biology*, 55(March 2018), 106–119. <https://doi.org/10.1016/j.semcancer.2018.04.001>

- Zhang, Y., Chen, M., Liu, C., Chen, J., Luo, X., Xue, Y., Liang, Q., Zhou, L., Tao, Y., Li, M., Wang, D., Zhou, J., & Wang, J. (2021). Sensitive and rapid on-site detection of SARS-CoV-2 using a gold nanoparticle-based high-throughput platform coupled with CRISPR/Cas12-assisted RT-LAMP. *Sensors and Actuators, B: Chemical*, 345. <https://doi.org/10.1016/j.snb.2021.130411>
- Zhang, Z. R., & Jiang, Z. R. (2022). Effective use of sequence information to predict CRISPR-Cas9 off-target. *Computational and Structural Biotechnology Journal*, 20, 650–661. <https://doi.org/10.1016/j.csbj.2022.01.006>

Appendix X
















Turnitin Similarity Report

Search or start new chat

Yakın Doğu Üniversitesi

QUICK SUBMIT | NOW VIEWING: ALL PAPERS

Submit

<input type="checkbox"/>	AUTHOR	TITLE	SIMILARITY	FILE	PAPER ID	DATE
<input type="checkbox"/>	A Eyad	Abstract	0% 		2294572303	14-Feb-2024
<input type="checkbox"/>	A Eyad	Conclusion	0% 		2294572349	14-Feb-2024
<input type="checkbox"/>	A Eyad	CH I	6% 		2294572311	14-Feb-2024
<input type="checkbox"/>	A Eyad	CH II	9% 		2294572320	14-Feb-2024
<input type="checkbox"/>	A Eyad	CH III	13% 		2294572328	14-Feb-2024
<input type="checkbox"/>	A Eyad	CH IV	15% 		2294572339	14-Feb-2024
<input type="checkbox"/>	A Eyad	 Full Thesis	15% 		2294572358	14-Feb-2024

Fulya Yonucu Yesterday

You Appendix X.docx

