

Yalçın Han ERBAŞI

A Master's Thesis

AGU 2022

CLASSIFICATION OF MICRORNA-  
DISEASE ASSOCIATION AND  
MICRORNA-SPECIES ASSOCIATION  
BASED ON  
K-MER SEQUENCE REPRESENTATION

A THESIS  
SUBMITTED TO THE DEPARTMENT OF  
ELECTRICAL AND COMPUTER ENGINEERING AND  
THE GRADUATE SCHOOL OF ENGINEERING AND  
SCIENCE OF ABDULLAH GUL UNIVERSITY  
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF  
MASTER OF SCIENCE

By  
Yalçın Han Erbaşı  
May 2022

CLASSIFICATION OF MICRORNA-DISEASE  
ASSOCIATION AND MICRORNA-SPECIES  
ASSOCIATION BASED ON  
K-MER SEQUENCE REPRESENTATION

A THESIS  
SUBMITTED TO THE DEPARTMENT OF  
ELECTRICAL AND COMPUTER ENGINEERING AND  
THE GRADUATE SCHOOL OF ENGINEERING AND  
SCIENCE OF ABDULLAH GUL UNIVERSITY  
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF  
MASTER OF SCIENCE

By  
Yalçın Han ERBAŞI  
May 2022

## **SCIENTIFIC ETHICS COMPLIANCE**

I hereby declare that all information in this document has been obtained in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Yalçın Han ERBAŞI



## REGULATORY COMPLIANCE

M.Sc. thesis titled **Classification of microRNA-Disease Association and microRNA-Species Association Based on k-mer Sequence Representation** has been prepared in accordance with the Thesis Writing Guidelines of the Abdullah Gül University, Graduate School of Engineering & Science.

Prepared By

Yaçın Han ERBAŞI

Advisor

Assistant Professor  
Burcu BAKIR GÜNGÖR

Head of the Electrical and Computer Engineering Graduate Program

Associate Professor

Kutay İÇÖZ

## ACCEPTANCE AND APPROVAL

M.Sc. thesis titled **Classification of microRNA-Disease Association and microRNA-Species Association Based on k-mer Sequence Representation** and prepared by Yalçın Han ERBAŞI has been accepted by the jury in the Electrical and Computer Engineering Graduate Program at Abdullah Gül University, Graduate School of Engineering & Science.

31/05/2022

(Thesis Defense Exam Date)

### JURY:

Advisor : Assistant Professor Burcu BAKIR GÜNGÖR

Member : Assistant Professor M. Gökhan BAKAL

Member : Assistant Professor Özkan Ufuk NALBANTOĞLU

### APPROVAL:

The acceptance of this M.Sc. thesis has been approved by the decision of the Abdullah Gül University, Graduate School of Engineering & Science, Executive Board dated ..... /..... / ..... and numbered .....

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

**(Date)**

Graduate School Dean  
Prof. Dr. İrfan ALAN

ABSTRACT

CLASSIFICATION OF MICRORNA-DISEASE  
ASSOCIATIONS AND  
MICRORNA-SPECIES ASSOCIATIONS BASED ON  
K-MER SEQUENCE REPRESENTATION

Yalçın Han ERBAŞI  
MSc. in Electrical and Computer Engineering  
Advisor: Assistant Professor Burcu BAKIR GÜNGÖR  
May 2022

The dysregulated gene expression brings about a variety of diseases, and dysregulation of microRNA (miRNA) has a wide impact on disease development and cellular physiology. Thus, miRNAs play important roles in a variety of fundamental and significant biological processes related to human diseases. There are a lot of research about changes in the function of miRNAs have been published in many human diseases. Computational methods serve as a complementary process to traditional wet-lab experiments, which require many resources and time in terms of detecting potential miRNA-Disease associations. Furthermore, there is a need to present a novel approach that allows assignment of an unknown miRNA to its most likely species. An easy way to filter new data would be to ensure that the new miRNA is classified below the maximum distance to the species known to originate from.

In this thesis, a computational model has been proposed for identifying miRNA-disease and miRNA-Species associations by depicting the miRNAs with their k-mer sequence representation and by utilizing machine learning methodologies. The difference of our approach is which we reveal disease and species associated the sequences of miRNA store information. This put a question about the miRNA's chemical compounds and their associations with different types of species and diseases.

With this study, the new disease-disease and species-Species associations disclosed can be calculated for many different species and diseases, these approaches can develop to species and disease classification. Lastly, our study may open a door to redefine species and diseases classifications which have been used nowadays, also it may provide the improvement of treatment strategies and early diagnosis.

*Keywords: microRNA, miRNA-Disease association, miRNA-Species association, k-mer representation*

## ÖZET

# K-MER SEKANS GÖSTERİMINE DAYALI MICRORNA-HASTALIK İLİŞKİLERİNİN VE MICRORNA-TÜR İLİŞKİLERİNİN SINIFLANDIRILMASI

Yalçın Han ERBAŞI

Elektrik ve Bilgisayar Mühendisliği Bölümü Yüksek Lisans

Tez Yöneticisi: Dr. Öğr. Üyesi Burcu BAKIR GÜNGÖR

Mayıs 2021

Düzensiz gen ekspresyonu, çeşitli hastalıkları beraberinde getirir ve mikroRNA'nın (miRNA) düzensizliğinin, hastalık gelişimi ve hücrel fizyoloji üzerinde geniş bir etkisi vardır. Bu nedenle miRNA'lar, insan hastalıklarıyla ilgili çeşitli temel ve önemli biyolojik süreçlerde önemli roller oynar. Birçok insan hastalığında miRNA'ların işlevindeki değişiklikler hakkında birçok araştırma yayınlanmıştır. Hesaplamalı yöntemler, potansiyel miRNA-hastalık ilişkilerini saptamak açısından birçok kaynak ve zaman gerektiren geleneksel ıslak laboratuvar deneylerini tamamlayıcı bir süreç olarak hizmet eder. Ayrıca, bilinmeyen bir miRNA'nın en olası türlerine atanmasına izin veren yeni bir yaklaşım sunmaya ihtiyaç vardır. Yeni verileri filtrelemenin kolay bir yolu, yeni miRNA'nın köken aldığı bilinen türlere olan maksimum mesafenin altında sınıflandırılmasını sağlamak olabilir.

Bu tezde, miRNA'ları k-mer sekans gösterimleri ile betimleyerek ve makine öğrenme metodolojilerini kullanarak miRNA-hastalığı ve miRNA-tür ilişkilerini tanımlamak için bir hesaplama modeli önerilmiştir. Yaklaşımımızın farklılıkları, miRNA depolama bilgilerinin dizileriyle ilişkili hastalıkları ve türleri ortaya çıkarmamızdır. Bu, miRNA'ların kimyasal bileşikleri ve bunların farklı türler ve hastalıklarla ilişkileri hakkında bir soru ortaya koyar.

Bu çalışma ile ortaya çıkan yeni hastalık-hastalık ve tür-tür ilişkileri birçok farklı tür ve hastalık için hesaplanabilmekte, bu yaklaşımlar tür ve hastalık sınıflandırmasını geliştirebilmektedir. Son olarak, çalışmamız günümüzde kullanılan tür ve hastalık sınıflandırmalarının yeniden tanımlanmasına kapı aralayabilir, ayrıca tedavi stratejilerinin geliştirilmesini ve erken teşhis edilmesini de sağlayabilir.

*Anahtar kelimeler: microRNA, miRNA-hastalık ilişkisi, miRNA-tür ilişkisi, k-mer gösterimi*

# Acknowledgements

I would like to extend my thanks to my advisor Asst. Prof. Burcu Bakır G ng r for her supervision and supports. I also want to thank my family and my friends for their patience and tolerance.





# TABLE OF CONTENTS

<b>1. INTRODUCTION .....</b>	<b>13</b>
1.1 PROBLEM STATEMENT .....	13
1.2 THESIS ORGANIZATIONS .....	14
<b>2. BACKGROUND.....</b>	<b>15</b>
2.1 BIOLOGICAL BACKGROUND .....	15
2.1.1 <i>Discovery of MicroRNA</i> .....	15
2.1.2 <i>Biogenesis of MicroRNA</i> .....	15
2.1.3 <i>Mechanism of Action of MicroRNAs</i> .....	17
2.1.4 <i>The Role of MicroRNAs</i> .....	18
2.1.5 <i>The Nomenclature of miRNAs</i> .....	18
2.1.6 <i>miRNAs Databases</i> .....	18
2.2 MACHINE LEARNING.....	20
2.2.1 <i>Supervised Learning</i> .....	21
2.2.2 <i>Unsupervised Learning</i> .....	21
2.3 ELIMINATION OF DATA .....	22
2.3.1 <i>Clustering Algorithm</i> .....	22
2.4 FEATURE EXTRACTIONS.....	23
2.4.1 <i>K-mer Frequencies</i> .....	23
2.5 ARTIFICIAL DATA GENERATION .....	24
2.5.1 <i>Synthetic Minority Over-sampling Technique</i> .....	24
2.6 CLASSIFICATION METHODS.....	25
2.6.1 <i>Logistic Regression (LR)</i> .....	25
2.6.2 <i>k-Nearest Neighbors (kNN)</i> .....	26
2.6.3 <i>Naive Bayes (NB)</i> .....	27
2.6.4 <i>Linear Discriminant Analysis (LDA)</i> .....	28
2.6.5 <i>Random Forest (RF)</i> .....	28
2.6.6 <i>Decision Tree (DT)</i> .....	29
2.6.7 <i>Multilayer Perceptron (MLP)</i> .....	30
2.7 CLASSIFICATION MODEL VALIDATION .....	32
2.7.1 <i>Cross Validation</i> .....	32
2.8 CLASSIFICATION EVALUATION METRICS .....	32
2.8.1 <i>Logistic Loss Function</i> .....	33
2.9 HIERARCHICAL CLUSTERING METHODS.....	33
2.9.1 <i>Single Linkage Clustering</i> .....	34
2.9.2 <i>Complete Linkage Clustering</i> .....	34
2.9.3 <i>Average Linkage Clustering</i> .....	34
2.9.4 <i>Ward Clustering</i> .....	35
2.9.5 <i>Centroid Clustering</i> .....	35
2.10 DISTANCE METRICS .....	35
2.10.1 <i>Euclidean Distance</i> .....	36
2.10.2 <i>Canberra Distance</i> .....	36
2.10.3 <i>Manhattan Distance</i> .....	36
2.10.4 <i>Minkowski Distance</i> .....	36
2.11 COPHENETIC CORRELATION COEFFICIENT .....	37

<b>3. MATERIALS AND METHODS.....</b>	<b>38</b>
3.1 INPUT DATA .....	38
3.1.1 Dataset of miRNA-Species.....	38
3.1.2 Dataset of miRNA-Disease.....	38
3.2 DATA CLEANING .....	39
3.3 FEATURES .....	40
3.4 SAMPLING METHODOLOGY .....	40
3.5 CHOICE OF THE CLASSIFIER.....	41
3.6 POSTPROCESSING.....	41
3.7 CHOICE OF THE HIERARCHICAL CLUSTERING .....	42
<b>4. RESULTS AND DISCUSSIONS.....</b>	<b>44</b>
4.1 MULTI CLASS CLASSIFICATION .....	44
4.1.1 miRNA-Disease Association.....	44
4.1.2 miRNA-Species Association.....	46
4.2 HIERARCHICAL CLUSTERING.....	48
4.1.1 miRNA-Disease Association.....	48
4.1.2 miRNA-Species Association.....	50
<b>5. CONCLUSIONS AND FUTURE PROSPECTS .....</b>	<b>53</b>
5.1 CONCLUSIONS .....	53
5.2 SOCIETAL IMPACT AND CONTRIBUTION TO GLOBAL SUSTAINABILITY .....	54
5.3 FUTURE PROSPECTS .....	55

# LIST OF FIGURES

Figure 2.1 The steps of miRNA biogenesis .....	16
Figure 2.2 The Binding of miRNA to the 3' UTR region of target mRNA .....	18
Figure 2.3 A screenshot of human pre-miRNA hsa-mir-1-1 in the miRBASE database	19
Figure 2.4 A screenshot of human pre-miRNA hsa-mir-1-1 in the HMDD database ...	20
Figure 2.5 Cluster diagram generated using the UCLUST algorithm.....	23
Figure 2.6 Determining the class according to the nearest 3 and 6 neighborhoods.....	27
Figure 2.7 An example of decision tree.....	30
Figure 2.8 Example MLP model of a multi-classification ANN .....	31
Figure 3.1 An example of transformation from miRNA sequences to k-mer frequency feature vectors .....	40
Figure 4.1 Dendrogram graph with Centroid clustering method, Euclidean distance metrics in the miRNA-Disease data .....	49
Figure 4.2 Correlation graph between diseases .....	50
Figure 4.3 Dendrogram graph with Average clustering method, Canberra distance metrics in the miRNA-Disease data .....	51
Figure 4.4 Correlation graph between species .....	52

# LIST OF TABLES

Table 3.1 A Part of MP-Vector.....	42
Table 4.2 LogLoss value of multi classification algorithms trained on cleaned miRNA-Disease data.....	45
Table 4.3 LogLoss value of multi classification algorithms trained on over-sampled miRNA-Disease data .....	46
Table 4.4 LogLoss value of multi classification algorithms trained on cleaned miRNA-Species data.....	47
Table 4.5 LogLoss value of multi classification algorithms trained on over-sampled miRNA-Species data .....	47
Table 4.6 Comparison of clustering methods and distance metrics in the miRNA-Disease data.....	48
Table 4.7 Comparison of clustering methods and distance metrics on the miRNA-Species data.....	50

# LIST OF ABBREVIATIONS

DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
miRNA	MicroRNA
mRNA	Messenger RNA
pri-miRNA	Primary MicroRNA
pre-miRNA	Precursor MicroRNA
PKI	Public Key Infrastructure
PQC	Post-Quantum Cryptography
SG	Smart Grid
SMC	Secure Multiparty Computation
SMOTE	Synthetic Minority Over-sampling Technique
LR	Logistic Regression
kNN	k-Nearest Neighbors
NB	Naive Bayes
LDA	Linear Discriminant Analysis
PCA	Principal Component Analysis
SVM	Support Vector Machine
RF	Random Forest
DT	Decision Tree
CART	Classification and Regression Trees
MARS	Multivariate Adaptive Regression Splines
MLP	Multilayer Perceptron
ANN	Artificial Neural Networks
LogLoss	Logistic Loss Function

# Chapter 1

## Introduction

### 1.1 Problem Statement

The branches of science that have been studied by humanity since ancient times have been named as fundamental sciences and the problems encountered have been classified among these disciplines and solutions have been sought. In time, problems that a few of the fundamental sciences can find solutions together have emerged and interdisciplinary work has gained importance. With the increase in the subjects that need to be studied the different disciplines, new branches of science have emerged. One of these new branches is bioinformatics that includes various branches of biology, more particularly molecular biology and computer technology and related data processing devices. In other words, it is the science of compiling and analyzing complex biological data.

Developments on gene and genome research, detection of protein three-dimensional structures, and discovery of DNA sequences can be described as exciting developments for researchers working on this subject. Following the discovery of DNA and RNA sequences, obtaining their sequences, and then evaluating these sequences in the light of statistical and computer science methods has created a new scientific horizon. In the meantime, there have been developments on the biological side as well. The MicroRNA (miRNA) molecules, whose existence was first discovered in 1993 and only named in 2001, have become the favorite of many researchers working in the field of bioinformatics. While it was not known exactly what its function was in the cell at the beginning, today it is known what function certain types of miRNAs have in the cell.

Classification is a method used to reveal hidden patterns in the database. By classification, the database is divided into small homogeneous groups according to certain characteristics. Classification is an analysis technique that shows which class a new data belongs to and is based on a learning algorithm. The purpose of these algorithms is to create a classification model and to determine a class for a data whose class is unknown. Well-defined features play a key role here. For this reason, various numerical mapping techniques are used to digitize a genetic sequence, that is, to generate features. In this thesis study, k-mer representation from these mapping techniques was used to create a classification model.

## **1.2 Thesis Organizations**

In the next chapter, information is given about the protein synthesis process, the biological interaction of tRNA and mRNA, and the formation of miRNA. In the same chapter, classification algorithms and clustering methods and the measurement metrics used are examined. In the third chapter, the problem that forms the basis of this thesis is presented and the solution presented to this problem is explained. Accordingly, the performance of classification and clustering methods used to predict a miRNA-disease and miRNA-Species associations are examined, while the methods and tools used in feature extraction, new sampling and data cleaning are described. Increasing the number of miRNAs causes an increase in the time required for the computation of features. For this reason, the use of a limited number of species and diseases has reduced the time spent for calculation. The developed method and data are explained in detail in the fourth section. Again, in this chapter, the experimental results obtained by testing the developed method on different data sets and the comparative interpretations of these results are given.

# Chapter 2

## Background

### 2.1 Biological Background

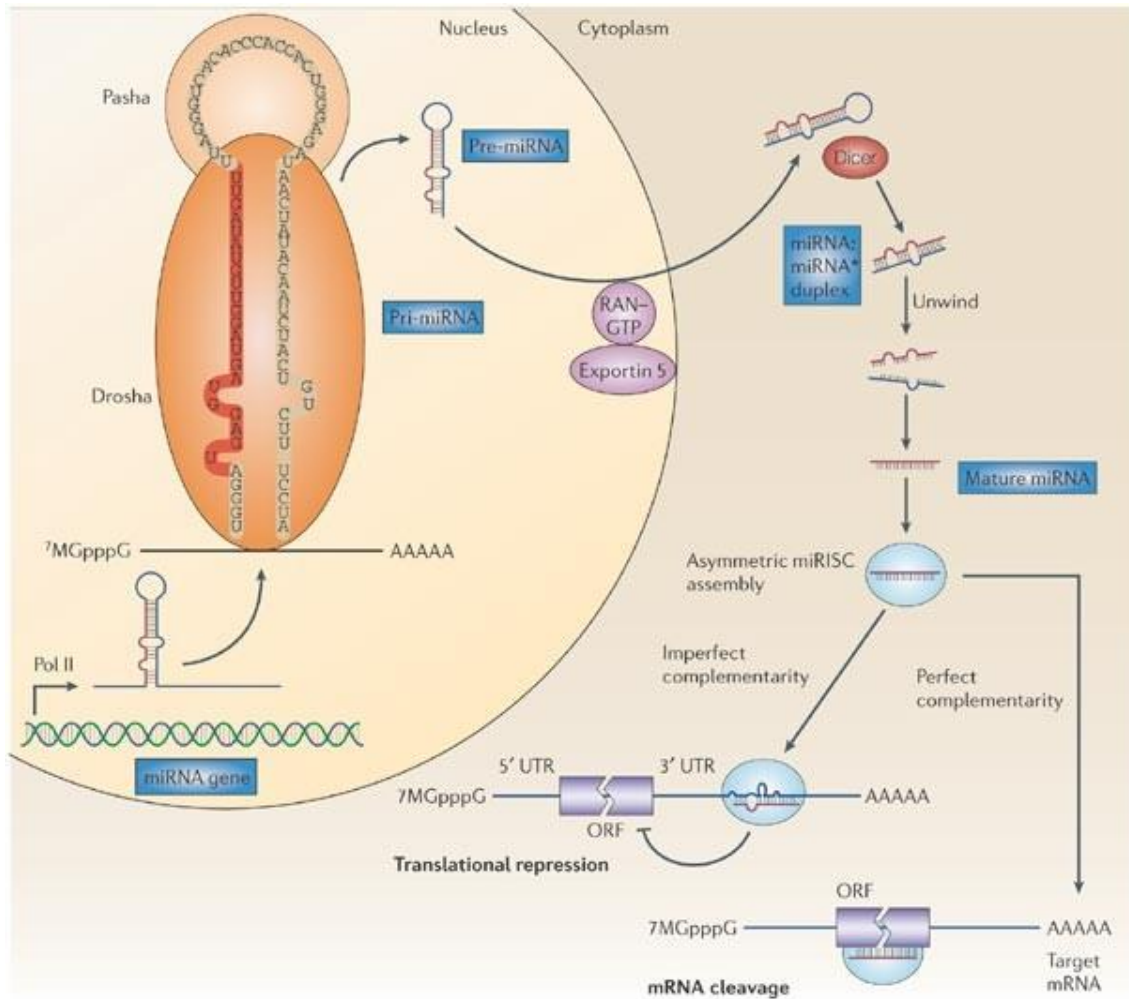
#### 2.1.1 Discovery of MicroRNA

MicroRNAs (miRNAs) are small, evolutionarily conserved non-coding RNAs about 18-25 nucleotides long. The first data on the existence of miRNAs began to be revealed in the mid-90s [1]. It was first discovered in the nematode *Caenorhabditis elegans* and its function was revealed in another laboratory. Combining the data of the two teams, they determined that the *lin-4* gene, which is important in the larval period, does not code for protein, but produces a small RNA product of 22 nucleotides, and this product suppresses the expression of the *lin-14* gene [1,2]. In 2000, it was observed that *let-7*, one of the genes controlling the developmental timing of *Caenorhabditis elegans*, encodes an RNA of approximately 22 nucleotides and that the *lin-41* gene, which is effective in larval development, is complementary to two different regions in the 3'UTR region [3]. In later studies, it was understood that the *let-7* gene was conserved among different species, including humans, and it was revealed that miRNAs may also be important in other organisms [4].

#### 2.1.2 Biogenesis of MicroRNA

Synthesis of miRNAs takes place in 3 stages. Initially, miRNAs are transcribed from gene regions on DNA as primary miRNA (pri-miRNA). Then these synthesized pri-miRNA molecules are transported to precursor miRNAs (pre-miRNAs) in the nucleus and finally to the cytoplasm, where they turn into mature miRNAs [5]. The steps of miRNA biogenesis can be seen in the Figure 2.1





**Figure 2.1 The steps of miRNA biogenesis**

miRNAs are synthesized by RNA polymerase II enzyme as pri-miRNA. Pri-miRNA (500-3000 bases), cap and stem-loop structure with poly A tail. In the nucleus, pri-miRNA is converted to pre-miRNA, approximately 70 nucleotides in length, by Drosha, an endonuclease of the RNase III family of enzymes, and its cofactor Pasha (DGCR8). The complex formed by 12 Pasha, a double-stranded RNA-binding protein, with Drosha, a nuclease, is called the microprocessor complex [6].

The molecule, which becomes pre-miRNA, is transported from the nuclear membrane to the cytoplasm in a RANGTP-dependent manner via a transport receptor, Exportin 5 [7]. Pre-miRNAs that pass into the cytoplasm are cleaved by the Dicer enzyme, an endonuclease from the RNase III family of enzymes, and converted into a double-stranded miRNA with a length of 18-24 nucleotides [8]. While Dicer cuts the stem-loop of the miRNA duplex, it also initiates the formation of RNA-induced silencing complex [9] and only one of the miRNA chains formed after cutting is

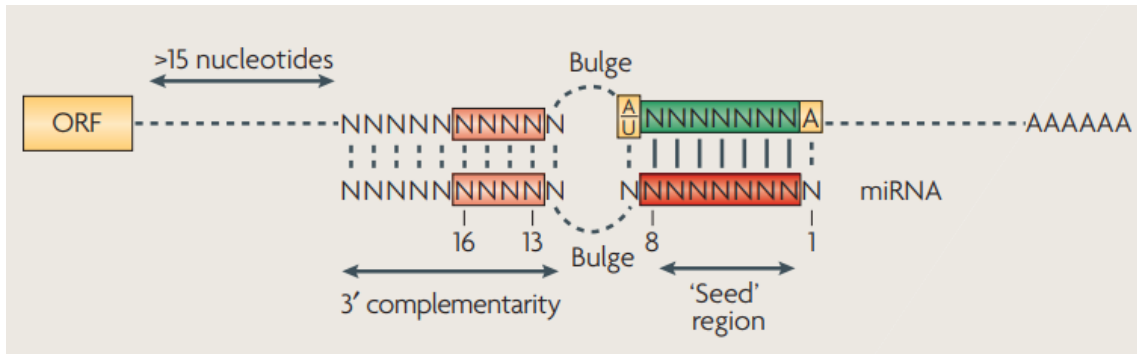
incorporated into the resulting RISC structure. With the effect of the Argonaute enzyme (RNase) (AGO) in the RISC structure, one of the two miRNA chains is selected (guide strand). The other chain is called anti-guide strand. The RISC complex digests as anti-guide thread substrate, also called passenger thread. miRNA molecules integrate with the RISC complex and act together with AGO proteins, causing suppression of translation (protein synthesis) or destruction of the mRNA molecule [10].

Both strands formed because of miRNA biogenesis are not equally efficient. Which of the formed chains will mature and become the active miRNA molecule is determined by the genetic code and thermodynamic interactions of the miRNA chain [11]. Studies with new generation sequencing methods have shown that the anti-guide strand expressed as miRNA\* is also not destroyed and can show activity [12].

### **2.1.3 Mechanism of Action of MicroRNAs**

Mature miRNAs are single-stranded structures with an average length of 20 nucleotides, non-coding RNA. They affect gene expression in the form of post-transcriptional gene regulation by binding to the 3'UTR, which is usually untranslated, and sometimes to the 5'UTR region of the target mRNA [13,14].

miRNAs can inhibit gene expression in two ways: mRNA degradation or inhibition of translation [13]. In cases where complete complementarity and binding to mRNA is observed, silencing of gene expression by mRNA cleavage is observed [15]. When mRNA fragmentation occurs, the cut portion is between residues 10 and 11 of the miRNA and the nucleotide pair. Even if the complementarity is not fully achieved, this situation does not change. The region between nucleotides 2-8 of miRNA is called the "seed" region. This region is a region where the connection with the mRNA is one-to-one, which also enables miRNA predictions to be made and can be seen in the Figure 2.2. [16]. This shows us that the cutting or cleavage point is not related to the base pair connection between the target and the miRNA, but rather to the residues in the miRNA, so it is the complementarity of the correct residues that matters. After mRNA degradation, the miRNA remains intact and can guide the destruction and recognition of additional messages [13].



**Figure 2.2 The Binding of miRNA to the 3' UTR region of target mRNA**

### 2.1.4 The Role of MicroRNAs

Studies give an idea about the role of miRNAs such as Lin-4 and Let-7 in controlling the timing of developmental transitions. However, this role is only one of the many functions of molecules in the cell. Studies conducted by inhibiting the Dicer enzyme and AGO protein, which have important roles during the synthesis (biogenesis) of miRNAs on many model organisms, show that these small RNA molecules take an active role in many 16 pathways such as embryo development, differentiation, proliferation, apoptosis, host-viral interactions and tumor-genesis [13]. This shows how important miRNA activity is for the biological processes of living things.

### 2.1.5 The Nomenclature of miRNAs


When nominating miRNAs, species names are placed at the front of the name with abbreviation. For example, the abbreviation “hsa” is used for human (homo sapiens). The use of the abbreviation “miR”, as in the example “hsa-miR-21”, indicates that it is a mature miRNA [17,18]. The mature miRNA, symbolized by “miR”, is formed from a stem-loop precursor sequence called pre-mir.

### 2.1.6 miRNAs Databases

The “miRBASE” database (<http://www.mirbase.org/>) was created in order to provide accurate and up-to-date information about miRNA to everyone. The first version (1.0) of this public database was published in December 2002 and contains 218 entries. In June 2014, it was generally updated and the 21st version was released and contains 28645 entries. As the last, it was updated again after a long break and the 22nd

version was started on March 2018 with 38589 entries. A screenshot of human pre-miRNA hsa-mir-1-1 in the miRBase database can be seen in the Figure 2.3 [19,20].

The screenshot displays the miRBase database entry for hsa-mir-1-1. The page header includes the miRBase logo and the University of Manchester logo. The navigation bar contains links for Home, Search, Browse, Help, Download, Blog, Submit, and a search box with the text 'hsa-mir-1-1'. The main content area is titled 'Stem-loop sequence hsa-mir-1-1' and contains the following information:

- Accession:** MI0000651 ([change log](#))
- Previous IDs:** hsa-mir-1b
- Symbol:** [HGNC:MIR1-1](#)
- Description:** *Homo sapiens* miR-1-1 stem-loop
- Gene family:** MIPF0000038; [mir-1](#)
- Literature search:**  [517 open access papers](#) mention hsa-mir-1-1 (2629 sentences)
- Stem-loop:**

```

5'   u   a                   gc   ---   a
     g g g a   a c a u a c u c u u u a u   c c a u a   u g g   c
      | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
3'   c u c u   u g u a u g a a g a a a u g u a   g g u a u   a u c   u
     a     a                   - a   c g a   g

```
- Deep sequencing:** [9922471](#) reads, [6.51e+03](#) reads per million, 159 experiments. Below this is a bar chart showing the distribution of reads across the sequence, with the sequence UGGGAACUACUCUUCUUUAUGCCCAUUGGACCGUCUAGCCUUAUGGAUGUAAGAGUGUGUCUCA.
- Confidence:** Annotation confidence: high. Feedback: Do you believe this miRNA is real?
- Comments:**

**Figure 2.3** A screenshot of human pre-miRNA hsa-mir-1-1 in the miRBASE database

Another Database is HMDD v3.2 which enable users to download, search and analyze supported miRNA-disease associations based on research. HMDD was first built in December 2007 and has been updated more than 30 times in the last 10 years. HMDD version 2 (HMDD v2.0) was released in June 2013. In version 2, data on miRNA disease were compiled in more detail. Data from genetics and epigenetics, data from circulating samples such as blood, plasma, and serum. HMDD v3.0 was released in June 2018 and now further categorized its evidence in the literature in latest version 3 in the Figure 2.4 [21,22].

**HMDD v3.2: the Human microRNA Disease Database version 3.2**

[Home](#)   [Browse](#)   [Search](#)   [miR-Target Network](#)   [Causality](#)   [Disease Network](#)  
[Download](#)   [Submit](#)   [Help](#)

You can submit a semicolon-delimited list of keywords(no more than 20) for batch search. [Use example](#)

[download the result file](#)

miRNA name	Evidence Code	Disease name	PMID	Description	Causality
hsa-mir-1-1	circulation_biomarker_diagnosis_down	Cardiomyopathy, Hypertrophic	<a href="#">17234972</a>	Significantly, the muscle-specific microRNA-1 (miR-1) was singularly downregulated as early as day 1, persisting through day 7, after aortic constriction. Induced hypertrophy in a mouse model.	NO
hsa-mir-1-1	circulation_biomarker_diagnosis_down	Cardiomyopathy, Hypertrophic	<a href="#">17468766</a>	downregulation	NO
hsa-mir-1-1	circulation_biomarker_diagnosis_down	Cardiomyopathy, Hypertrophic	<a href="#">17479098</a>	The researchers showed that expression of miR-1 and another muscle-specific miRNA, miR-133, is decreased in human and mouse hypertrophic heart tissue.	NO
hsa-mir-1-1	circulation_biomarker_diagnosis_down	Hypertrophy	<a href="#">21303526</a>	The downregulation of miR-1, miR-133a, and upregulation of miR-21 can be reversed by one single upstream regulator, SRF.	NO
hsa-mir-1-1	circulation_biomarker_diagnosis_ns	Hypertension	<a href="#">18690400</a>	miR-1: deregulation	NO
hsa-mir-1-1	circulation_biomarker_diagnosis_ns	Parkinson Disease	<a href="#">21295623</a>	miR-1, miR-22* and miR-29 expression levels allowed to distinguish non-treated PD from healthy subjects, miR-16-2*, miR-26a2* and miR30a differentiated treated from untreated patients.	NO
hsa-mir-1-1	circulation_biomarker_diagnosis_ns	Muscular Dystrophy, Duchenne	<a href="#">21425469</a>	miR-1, miR-133, and miR-206 are new and valuable biomarkers for the diagnosis of DMD and possibly also for monitoring the outcomes of therapeutic interventions in humans	NO
hsa-mir-1-1	circulation_biomarker_diagnosis_ns	Myocardial Infarction	<a href="#">23747779</a>	Circulating miR-1, miR-208a, and miR-133a continuously rose during the first 4 h after induction of AMI.	NO
hsa-mir-1-1	circulation_biomarker_diagnosis_up	Coronary Artery Disease	<a href="#">17919180</a>	elevated expression	NO
hsa-mir-1-1	circulation_biomarker_diagnosis_up	Cardiomyopathy, Hypertrophic	<a href="#">17965831</a>	miR-1 is overexpressed in patients with coronary artery disease and that overexpression of miR-1 in a rat model of cardiac infarction exacerbated arrhythmogenesis. Cardiac hypertrophy may also be regulated by miR-1, which is significantly down-regulated in hypertrophic tissue.	NO
hsa-mir-1-1	circulation_biomarker_diagnosis_up	Coronary Artery Disease	<a href="#">17965831</a>	miR-1 is overexpressed in patients with coronary artery disease and that overexpression of miR-1 in a rat model of cardiac infarction exacerbated arrhythmogenesis. Cardiac hypertrophy may also be regulated by miR-1, which is significantly down-regulated in hypertrophic tissue.	NO
hsa-mir-1-1	circulation_biomarker_diagnosis_up	Retinal Degeneration	<a href="#">18834879</a>	miR-1: upregulation	NO
hsa-mir-1-1	circulation_biomarker_diagnosis_up	Myocardial Infarction	<a href="#">19245789</a>	miR-1: Upregulated expression in a rat model	NO
hsa-mir-1-1	circulation_biomarker_diagnosis_up	Muscular Dystrophy, Duchenne	<a href="#">21479190</a>	the serum levels of several muscle-specific miRNAs (miR-1, miR-133a and miR-206) are increased	NO
hsa-mir-1-1	circulation_biomarker_diagnosis_up	Myocardial Infarction	<a href="#">21642241</a>	Increased MicroRNA-1 and MicroRNA-133a Levels in Serum of Patients With Cardiovascular Disease Indicate Myocardial Damage.	NO

**Figure 2.4** A screenshot of human pre-mirRNA hsa-mir-1-1 in the HMDD database

## 2.2 Machine Learning

Machine learning can be defined as giving machines the ability to make the perceptions that human intelligence can make by examining the data at hand and to use them as data for subsequent processes. Machine learning provides systems with the ability to learn automatically and evolve from experience. Also these methods are called cutting edge technologies in the fourth industrial revolution (Industry 4.0) [23]. The purpose of machine learning algorithms is to optimize the performance of a task using data or experience. In these data-driven algorithms, the performance of machine learning increases as the size of the data increases. This is similar to how a person can

perform a certain task better by gaining more experience. Machine learning systems work in two processes, learning (training) and testing. In the learning phase, the task is experienced with the training data and learning ends when the learning performance reaches a satisfactory point. The model developed through the training process can then be used for classification, clustering or regression [24]. Today, special machine learning algorithms have been developed that can solve complex problems. Machine learning algorithms are divided into two as supervised and unsupervised machine learning according to the structure of the data set used and the way they learn from the data set.

### **2.2.1 Supervised Learning**

Supervised learning is machine learning in which the relationship with the dependent variable ( $y$ ) is learned using independent variables ( $x$ ). The purpose of supervised learning is to estimate the output variables ( $y$ ) of this data set, in case a new data set is used, by estimating the relationship between the independent variables and the dependent variable at the best level. In supervised learning, an algorithm from the training dataset can be thought of as a teacher supervising the learning process. The algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm reaches an acceptable level of performance [25,26]. This learning method is divided into classification and regression. Since regression algorithms were not used in our study, detailed information about the algorithms will not be given.

### **2.2.2 Unsupervised Learning**

Unsupervised Machine Learning is machine learning with only independent variables and no dependent variables. The purpose of unsupervised learning is to model the underlying structure in the data to learn more about the data. In unsupervised learning, there are no data labels, that is, no supervision. The goal is to discover and present the structure in the data. Unsupervised learning is useful for clustering and association problems. K-means for clustering problems and a priori algorithm for relationship rule learning problems can be given as examples of unsupervised machine learning [25]. Clustering is an unsupervised learning task that aims to find hidden patterns in unlabeled input data in the form of clusters. It involves organizing data into meaningful natural groups based on similarity between different features to learn the

structure of the data. Clustering algorithms, which involve arranging the data to have high similarity within clusters and low similarity between clusters, are used in many applications in the fields of machine learning, data mining, network analysis, pattern recognition and computer vision [27].

## **2.3 Elimination of Data**

Cleaning the data is usually the initial step in any machine learning approach. The goal of data cleaning is to find and eliminate mistakes as well as duplication in order to build a trustworthy dataset. This enhances the training data's quality.

To find duplication, one must be able to compare two sequences, which is accomplished by alignment. Because there are several alternative alignments, one must choose the best one. The optimal alignment is one in which the two sequences are matched so that the majority of nucleotides match.

BLAST is a technique for comparing protein and nucleotide sequences (of which miRNAs are a part) [28]. BLAST may compare sequences to an entire database or a different group of sequences. Because two sequences might be quite close even if they are not identical and differ just by a few nucleotides, BLAST yields a significance in how similar they are.

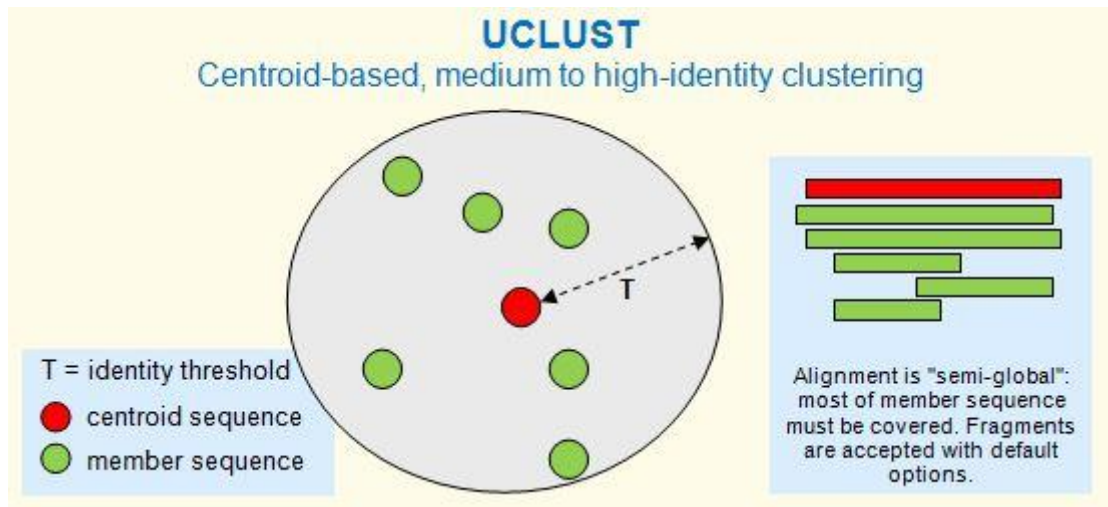
Local alignments are used by BLAST, which means it compares parts of sequences to each other. Of course, how the two sequences are aligned to one other has a role. As a result, BLAST attempts to identify the best alignment between two sequences. It accomplishes this by attempting to match brief segments of the sequence in question to other sequences first. The high-matching sequences are then lengthened in order to identify better and longer matching sequences.

However, if one wishes to utilize BLAST to eliminate duplicates, extra processing is required after receiving the results because BLAST only provides related sequences and their relevance. However, it's also necessary to choose which of these sequences to preserve and which to discard.

### **2.3.1 Clustering Algorithm**

As a result, sequences are frequently grouped. Usearch is a clustering technique that may be used to eliminate duplicates in miRNA datasets. Usearch has a clusterer that clusters data using the UCLUST algorithm. The cluster representatives may then be

utilized as the cleaned dataset in the Figure 2.45 [29]. Fortunately, the Usearch toolbox has all of those stages already implemented, and the final dataset can be acquired with a single command.



**Figure 2.5 Cluster diagram generated using the UCLUST algorithm**

## 2.4 Feature Extractions

Every classification approach necessitates the inclusion of some features that may be employed in the classification process. Because miRNA is employed in this research, certain information from the miRNA sequence must be retrieved so that the classifier can use it. It is critical to choose appropriate features since they might have a significant influence on the classifier's accuracy [30]. Even though the chosen classification technique is appropriate in the case, if the characteristics are insufficient to distinguish between the various groups, for example because there is no association between the feature and the groupings, the outcome may be bad.

### 2.4.1 K-mer Frequencies

K-mers are nucleotide sequences having a length of  $k$ . A 1-mer across the letters A, C, T, G, for example, can yield the sequences A, C, T and G. The sequences AA, AC, AT, AG, CA, CC, CT, CG, TA, TC, TT, TG, GA, GC, GT and GG are produced by a 2-mer.

The  $k$ -mer frequency defines how frequently a certain  $k$ -mer appears in sequence in relation to its length. To compute the frequencies of all these  $k$ -mers in a sequence, just count how many times each  $k$ -mer appears in sequence in the question and divide



by the length of the sequence in question.  $S(i, j) = s_i \dots s_j$  is the subsequence of  $S$  from nucleotide  $i$  to nucleotide  $j$ , and the probability of a  $k$ -mer  $m$  with length  $k$  is:

$$m_{freq} = \frac{\sum_{i=0}^{n-k} \text{mer}(m, S(i, i+k))}{n} \quad (2.1)$$

where  $\text{mer}(m, S(i, j))$  is a function that returns one if the  $k$ -mer  $m$  equals to the subsequence  $S(i, j) = s_i \dots s_j$  and zero otherwise [31]:

$$\text{mer}(m, S(i, j)) = \begin{cases} 1 & \text{if } S(i, j) = m \\ 0 & \text{otherwise} \end{cases} \quad (2.2)$$

For instance,  $S$  represent the ACCATTGA sequence and  $m$  represent the 2-mer AT. The frequency of  $m$  in  $S$  is therefore  $2/8$ . Because AT appears twice in  $S$  and  $S$  is eight letters long.

## 2.5 Artificial Data Generation

In many classification problems, class imbalances must be dealt with, when one class has more samples to model than the other. These situations are particularly common when working with biological data, as in our study.

When classifiers are trained on imbalanced classes, they tend to favor the larger one since they were exposed to more examples of the larger class. There are numerous approaches of dealing with this issue. Researchers categorize the most prevalent techniques into algorithmic, sampling and feature selection [32].

The main class is either under-sampled, the minor class is over-sampled, or a combination of both is employed in the sampling strategy. Random samples are eliminated from the primary class when under-sampling is used. This might result in the loss of important data. When data is oversampled, false data is created from the original data.

### 2.5.1 Synthetic Minority Over-sampling Technique

Synthetic Minority Over-sampling Technique (SMOTE) is a popular tool for generating this type of data [33]. SMOTE deal with every sample for minor class, also

computes its k-nearest neighbors in the same class to produce fresh examples for the minor class. The default value for k in Smote is 5. A reasonable number of minor class samples are randomly picked based on the main and minor class differences. To interpolate between these locations and construct a new sample, the k-nearest neighbors are employed. The problem of overfitting is avoided in this way.

To cope with the problem of class imbalance, research adopted a k-medoid under-sampling strategy [32]. The k-medoid technique is similar to the k-means strategy (which is detailed later), but instead of averaging points, it utilizes medoids to represent clusters. A medoid is a cluster point with the shortest distance between it and all other cluster points. They utilized the k-medoid technique on the major class with k equal to the number of points in the minor class as the dataset for the major class [32]. This strategy, according to their research, considerably enhances the classifier's accuracy.

It is attempted to maximize the classifier performance for unseen data in the algorithm method. There are various methods for accomplishing this purpose. Using a single class classifier is one option. Because it is trained just on one class and ignore everything else, the imbalanced classes problem is avoided when employing one class classifiers. However, when working with high-dimensional data, these strategies may not be sufficient. As a result, while working with such data, feature selection is frequently utilized. j features are chosen from the feature set for feature selection, allowing the classifier to achieve maximum performance.

## **2.6 Classification Methods**

### **2.6.1 Logistic Regression (LR)**

Logistic regression is a supervised learning classification algorithm used to estimate the probability of a target variable. The nature of the target or dependent variable is binary, which means there will only be two possible classes.

Mathematically, a logistic regression model predicts  $P(Y=1)$  as a function of  $X$ ., It is one of the simplest machine learning algorithms that can be used for various classification problems such as spam detection, diabetes prediction, cancer detection.

Logistic regression generally means binary logistic regression with dual target variables, although there may be two other categories of target variables that can be

predicted by it. Based on this number of categories, Logistic regression can be divided into the following types:

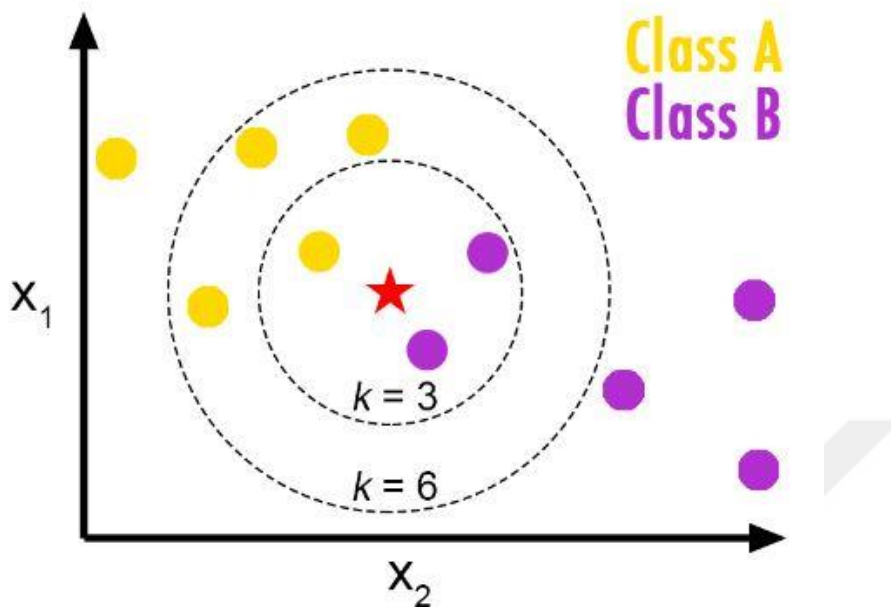
- **Binary or Binomial:** In this type of classification, a dependent variable will have only two possible types, 1 and 0. For example, these variables can represent success or failure, yes or no, winning or losing.
- **Multinomial:** In this type of classification, the dependent variable may have 3 or more possible unordered types or types without quantitative significance. For example, these variables could represent “Type A” or “Type B” or “Type C”.
- **Ordinal:** In this type of classification, the dependent variable can have 3 or more ordinal types or types with quantitative significance. For example, these variables may represent “poor” or “good”, “very good”, “excellent” and each category may have scores such as 0, 1, 2, 3.

In the case of binary logistic regression, the target variables should always be binary and the desired outcome is represented by factor level 1. There should be no multicollinearity in the model, that is, the independent variables should be independent of each other. Significant variables should be added to the model. A large sample size should have been chosen for logistic regression

### **2.6.2 k-Nearest Neighbors (kNN)**

The purpose of the algorithm is to classify a new object with the help of pre-classified examples by taking advantage of its properties. The algorithm searches the pattern space to determine which class an unknown sample belongs to and finds the  $k$  sample that is closest to the unknown sample. Then, the unknown sample is assigned to the class with which it is most similar out of the  $k$  nearest neighbors. One of the methods to be used to calculate the proximity is the Euclidean distance. In this algorithm, apart from Euclidean distance, Minkowski, Manhattan, Chebyshev and Dilca distances can also be used. The process flow of this algorithm is as follows; The parameter  $k$ , which expresses the number of nearest neighbors to a given point, is determined. The distances between this point and all other points in the data set are calculated. Then, according to these calculated distances, the rows are ordered from smallest to largest, and the smallest  $k$  of them is selected. Then, it is determined which class the selected rows belong to, and the most repetitive class is selected. The selected

class is accepted as the class of the observation value. An example of the nearest 3 and 6 neighbors is given in the Figure 2.6 [34].



**Figure 2.6 Determining the class according to the nearest 3 and 6 neighborhoods**

### **2.6.3 Naive Bayes (NB)**

Naive Bayes algorithm is based on Bayes theorem, which assumes independence between each pair of variables. Document or text classification, spam filtering, etc. can be used for both binary and multiclass categories [23]. This algorithm takes its name from the English mathematician Thomas Bayes, who lived in the 17th century. Naive Bayes is an algorithm that performs operations according to probability calculation. It processes the training data according to the probability formula and outputs a percentage for each condition. Test data performs classification according to the calculated probabilities. Naive Bayes classifier algorithm performs well when the arguments are categorical. It is also a suitable method for multi-class predictions. In Naive Bayes classification, a certain percentage of labeled data is transferred to the model. With the probability operations on the labeled data, the new test data given to the model is run according to the previously obtained probability values and it is tried to determine which class the given test data is. The greater the number of labeled data, the more likely it is to detect the true class of test data [26,35].

Bayes theorem shows the relationship between conditional probabilities and a priori (marginal) probabilities for a random variable.

$$P(A \setminus B) = \frac{P(B \setminus A)P(A)}{P(B)} \quad (2.3)$$

In Equation 2.3,  $P(A \setminus B)$ : the probability of event A occurring if event B occurs;  $P(B \setminus A)$ : The probability of event B occurring when event A occurs;  $P(A)$  ve  $P(B)$ : a priori probabilities of events A and B.

A naive bayes classification problem consists of many features and an outcome (target) variable.

$$P(C \setminus F_1, \dots, F_n) = \frac{P(F_1, \dots, F_n \setminus C)P(C)}{P(F_1, \dots, F_n)} \quad (2.4)$$

In Equation 2.4, C denotes the given target and F denotes the properties. The working principle of this algorithm is simply the product of all conditional probabilities.

#### **2.6.4 Linear Discriminant Analysis (LDA)**

Linear discriminant analysis (LDA) is a generalized form of Fisher's linear discriminant. LDA is a technique that can be applied to a variety of fields, including statistics, machine learning and pattern recognition. The final product can be used as a linear classifier. LDA makes a concerted effort to distinguish between two or more classes. The goal of LDA is to avoid overfitting and save time.

Linear separation analysis is closely related to principal component analysis (PCA) and factor analysis in that they examine the linear combination of variables in a data that best explains the data [36]. LDA finds a union that separates the given classes, while PCA ignores the classes. Factor analysis differs from LDA in that it examines variance rather than in-class similarity and models latent variables.

#### **2.6.5 Random Forest (RF)**

Random forest algorithm is an ensemble classification technique used in machine learning and data science in various application areas [23]. In this method, more than one decision tree is created, and these trees are combined to obtain a more accurate and

stable forecast. Each tree in the forest makes a class prediction, and the class with the most votes becomes the model's prediction. The Gini index, which is one of the most widely used techniques and used in decision trees, is used to determine the branching criteria in the random forest method. The operation of this algorithm is based on two different parameters, which are the number of variables used in each node and the number of trees to be developed, respectively. Random forest adds dependency on luck in addition to the model while growing trees. When dividing a node, instead of looking for the most important feature, it looks for the best feature within a random subset of features. This results in a wide variation, which usually results in a better model. One of the advantages of the random forest algorithm is that it is possible to measure the relative importance of each variable on its estimation. Another advantage is that it can be used for both classification and regression problems. The disadvantage of this method is that a large number of trees can make the algorithm very slow and ineffective for real-time predictions. In the Gini index, the frequency table of the variables is created by grouping each variable in pairs [37]. Equivalent for each variable in Equation 2.5 values are calculated.

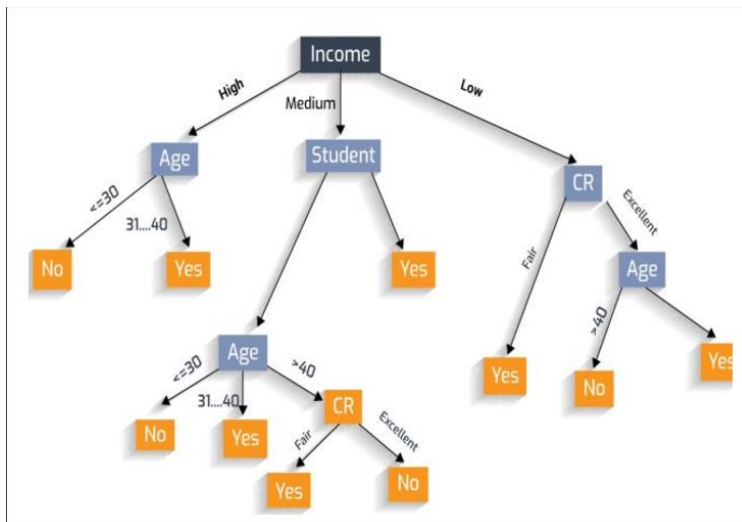
$$\text{Gini}(n) = 1 - \sum_{j=1}^2 (p_j)^2 \quad (2.5)$$

### 2.6.6 Decision Tree (DT)

Decision Tree is a classification algorithm used in many fields. Decision tree learning methods are used for both classification and regression tasks [23]. Algorithms such as CART (Classification and Regression Trees), ID3, C4.5, C5.0 and MARS (Multivariate Adaptive Regression Splines) are decision tree classification algorithms. Decision trees are preferred because they are easy to use and interpret. The structure of a decision tree includes root, node, branch and leaf. The lowest part is called the "leaf" and the upper part is called the "root". The variables in the dataset represent the nodes. The connection between the nodes is provided by structures called "branches". Deciding on which variable value to branch is the most important step in constructing a decision tree. In these methods, information gain, Gini index and Towing rule are widely used as decision making criteria.

CART algorithms are used as a decision tree algorithm. The CART algorithm, which is used for both classification and regression purposes, has used the Gini index

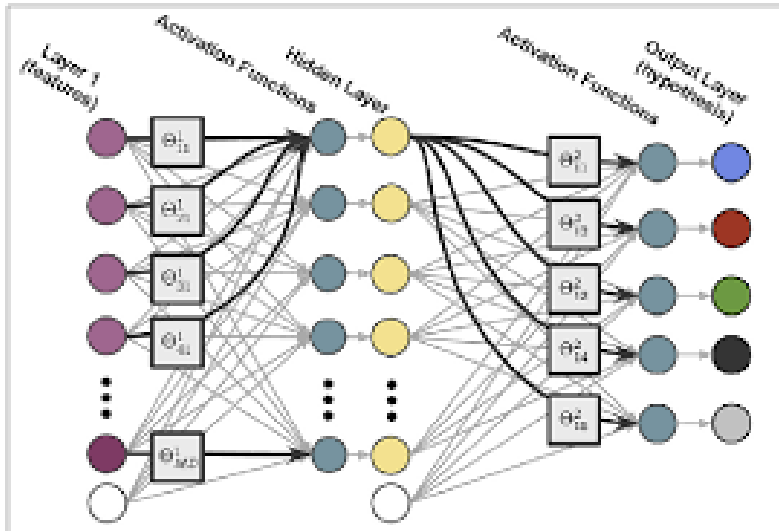
given in equation 2.5 as a branch. The CART algorithm ensures that the relevant group is divided into two more homogeneous subgroups at each step [38].



**Figure 2.7** An example of decision tree

### 2.6.7 Multilayer Perceptron (MLP)

Multilayer Perceptron is an Artificial Neural Networks (ANN) algorithms have been developed inspired by the learning process of the human brain. In artificial neural networks, artificial nerve cells are modeled to be interconnected, just as neurons in the biological nervous system interact with each other. It is thought that these models, called architecture or network, will have the capacity to reveal and learn the relationship between data [39]. The artificial nerve cell, inspired by the biological nerve cell, was initially developed as a single layer perceptron, which can only perform linear classification and is considered the most primitive artificial neural network. However, a single-layer perceptron is not sufficient for solving complex problems. For this reason, the multilayer perceptron shown in Figure 2.8 and used today has been developed [40].



**Figure 2.8 Example MLP model of a multi-classification ANN**

Multilayer sensors consist of input, hidden and output layers. Unlike single-layer sensors, these sensors can make nonlinear classifications. The number of hidden layers in these sensors and the number of neurons in these layers can vary [39]. In the artificial neural network architecture, the values of the connections connecting the neurons are called weights [41]. Neural networks perform iterative processing using the training dataset first to calculate weight estimates. It then applies these parameter estimates to the validation data to calculate the error values. The process then happens iteratively, passing parameter estimates from the adjusted training data to the validation dataset to find the smallest possible error relative to the validation data set. Therefore, the algorithm monitors the error associated with the validation dataset while using the training dataset to implement gradient descent. When the final weight estimates reach the smallest error in the validation data set, the most appropriate point in the process is determined [42]. The multilayer perceptron model works with the feed forward method and back propagation algorithm. The back propagation algorithm takes place in two stages, forward and backward. In the first step, the variables are presented to the training networks. A weight value from the normal distribution family is assigned for each neuron of each variable, and after the values are multiplied and summed with the weights, the bias value is added and an activation function is sent to all neurons in the next layer with an activation function in the neurons in the hidden layer [41].



## **2.7 Classification Model Validation**

### **2.7.1 Cross Validation**

The challenge with machine learning is that the performance of the classifier is constantly dependent on the data which was trained on. Even if only a little amount of data is given, the goal is to build a classifier that can generalize across the entire data space. As a result, it's critical to select the training set in such a way that this is attainable. To gain a representative training set, random selection methods are usually used. However, because to the nature of the samples, the classifier's performance can vary even with this strategy. As a result, cross validation is frequently employed. Typically, the dataset is partitioned into a test and training set at random. The classifier is then trained and tested on the training and test sets. The goal of cross validation is to produce a more dependable result by repeating the process numerous times, with each iteration's training and testing sets being picked at random. The average of all iterations is usually the final classifier score.

The most prevalent type of cross validation is K-fold cross validation [43]. Instead of selecting training and testing data at random in each run, the data is randomly divided into k equally sized "folds" up front in this approach. The test set is one of these k folds, whereas the training set is the remainder. Another fold will be utilized as the training set in the next run, while the remaining k-1 sets will be used as the test set. This is done until every fold has been through the training set at least once.

## **2.8 Classification Evaluation Metrics**

Model success obtained by constructing a model or using existing models to solve the classification problem can be thought of as the number of correct predictions from all predictions made. However, this considered information only gives the accuracy information of the classification. Accuracy value will not be enough to compare multi classification methods. Therefore, Logistic Loss evaluation which is consistent for multi-classification problems that is used in this study.

### 2.8.1 Logistic Loss Function

Logistic Loss (LogLoss) is an important criterion for classification based on predictive probability values. It takes into account the uncertainty of the estimate based on how much it varies from the true value. This gives a more consistent perspective on the performance of our model.

LogLoss is an important criterion for classification based on predictive probability values. It measures the performance of a classification model with an output of a probability value between 0 and 1, considering the uncertainty of how much the predicted value differs from the true value. The aim of model is to minimize log loss value. A perfect model would have a LogLoss of 0.

$$\text{logloss} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M y_{ij} \log(p_{ij}) \quad (2.6)$$

N = the number of samples in the test set

M = the number of class labels

log = the natural logarithm

y = 1 if observation i is in class j and 0 otherwise

p = the predicted probability that observation i belongs to class j

$$\max(\min(p, 1 - 10^{-15}), 10^{-15}) \quad (2.7)$$

To avoid the extremes of the LogLoss function, predicted probabilities are replaced with Equation 2.7.

## 2.9 Hierarchical Clustering Methods

Cluster analysis is one of the most important data mining processes to group variables according to their similarities and to obtain summary information about objects belonging to the same group through these groups. Therefore, it is not known how many clusters will be formed as a result of the initial phase and which attributes will affect this clustering process.

Clustering methods are examined under two headings: hierarchical and non-hierarchical. In this study, hierarchical clustering were implemented to creates a cluster by calculating the relationship between the data in the distance or similarity matrices.

### 2.9.1 Single Linkage Clustering

It is defined as the minimum distance between two clusters. It does not take into account the cluster structure. It is also called the nearest neighbor [44].

The minimum distance between the two clusters  $C_1$  and  $C_2 \cup C_3$  is calculated as given in Equation 2.8.

$$d(C_1, C_2 \cup C_3) = \min[d(C_1, C_2), (C_1, C_3)] \quad (2.8)$$

d=distance between two clusters

### 2.9.2 Complete Linkage Clustering

It is defined as the maximum distance between two clusters. It does not take into account the cluster structure like the single link method. It is also called the furthest neighborhood [45].

The maximum distance between two clusters  $C_1$  and  $C_2 \cup C_3$  is calculated as given in Equation 2.9.

$$d(C_1, C_2 \cup C_3) = \max[d(C_1, C_2), (C_1, C_3)] \quad (2.9)$$

d = distance between two clusters

### 2.9.3 Average Linkage Clustering

The distance between two clusters is the average of the distance between all data pairs consisting of one sample from each group. It is also accepted as the unweighted pair group method using the mean connection approach [49,50] The average link aggregation method is calculated as given in Equation 2.10.

$$d(C_1, C_2 \cup C_3) = \frac{n_2 \cdot d(C_1, C_2) + n_3 \cdot d(C_1, C_3)}{n_2 + n_3} \quad (2.10)$$

Here  $n_1$ ,  $n_2$  and  $n_3$  are sample data pairs in sets  $C_1$  and  $C_2$ , respectively.

d = distance between two clusters

n = number of data

## 2.9.4 Ward Clustering

The Ward method obtains new clusters by minimizing the intra-cluster variance. Among these clusters, it chooses the cluster with the lowest square error value[47]. Ward clustering method is calculated as given in Equation 2.11.

$$d = \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} \quad (2.11)$$

d = distance between two clusters

x= i'th observation

n = number of data

## 2.9.5 Centroid Clustering

The Ward method obtains new clusters by minimizing the intra-cluster variance. Among these clusters, it chooses the cluster with the lowest square error value[47]. Ward clustering method is calculated as given in Equation 2.12.

$$d(C_1, C_2 \cup C_3) = \frac{n_2}{n_2+n_3} d(C_1, C_2) + \frac{n_3}{n_2+n_3} d(C_1, C_3) - \frac{n_2 n_3}{(n_2 + n_3)^2} d(C_2, C_3) \quad (2.12)$$

d = distance between two clusters

n = number of data

## 2.10 Distance Metrics

The distance between two units is less than or equal to the sum of the distances of these two units to a third unit:

- Positivity  
 $d(i, j) \geq 0$
- Reflection  
 $d(i, j) = 0 \iff i \leftrightarrow j$
- Symmetry  
 $d(i, j) = d(j, i)$
- Triangle inequality  
 $d(i, j) \leq d(i, k) + d(k, j)$

### 2.10.1 Euclidean Distance

Euclidean Distance is one of the popular and classical similarity measures used in clustering methods. Euclidean distance is defined as the distance between two points or vectors [48]. The Euclidean Distance is calculated as given in Equation 2.13.

$$d_{\text{Euclidean}}(T_1, T_2) = \sum_{j=1}^n \sqrt{(T_{1j} - T_{2j})^2} \quad (2.13)$$

( $T_1$ ) and ( $T_2$ ) = Two points or vectors

### 2.10.2 Canberra Distance

It is a sensitive distance measure for small points with non-negative values close to zero [49]. The Canberra Distance is calculated as given in Equation 2.14.

$$d_{\text{Canberra}}(x_i, x_j) = \sum_{l=1}^d \frac{|x_{il} - x_{jl}|}{|x_{il}| + |x_{jl}|} \quad (2.14)$$

( $x_i$ ) and ( $x_j$ ) = Two points or vectors

### 2.10.3 Manhattan Distance

The Manhattan distance between two points is expressed as the sum of the absolute differences of their coordinates [50]. Manhattan Distance is calculated as given in Equation 2.15.

$$d_{\text{Manhattan}}(x_i, x_j) = \sum_{l=1}^d |x_{il} - x_{jl}| \quad (2.15)$$

( $T_1$ ) and ( $T_2$ ) = Two points or vectors

### 2.10.4 Minkowski Distance

The Minkowski distance is defined as a metric in vector space, which can be considered a generalization of both the Euclidean distance and the Manhattan distance [48]. The Minkowski Distance is calculated as given in Equation 2.16.

$$d_{\text{Minkowski}}(T_1, T_2) = \left( \sum_{i=1}^n |T_{1i} - T_{2i}|^p \right)^{\frac{1}{p}} \quad (2.16)$$

( $T_1$ ) and ( $T_2$ ) = Two points or vectors

## 2.11 Cophenetic Correlation Coefficient

The cophenetic correlation coefficient is a coefficient calculated to evaluate the agreement between the raw data distances and the distance measures used [55,56] It is widely preferred to evaluate both an appropriate distance criterion of dataset classification and the efficiency of various clustering techniques [50,57]. The high cophenetic correlation coefficient indicates that it is the most accurate clustering and distance criterion for the data set [55,56]. The cophenetic correlation coefficient is calculated as given in Equation 2.17.

$$C = \frac{\sum_{i < j} (x(i, j) - \bar{x})(t(i, j) - \bar{t})}{\sqrt{\sum_{i < j} [x(i, j) - \bar{x}]^2 \sum_{i < j} [t(i, j) - \bar{t}]^2}} \quad (2.17)$$

$x(i, j) = |X_i - X_j| =$  Euclidean distance

$t(i, j) = |T_i - T_j| =$  Dendrogram distance

# Chapter 3

## Materials and Methods

### 3.1 Input Data

#### 3.1.1 Dataset of miRNA-Species

For numerous organisms such as viruses, bacteria, and eukaryotes, the miRbase database contains all miRNAs found to date. MiRBase which is formerly known as microRNA Registry version 21 provided whole of miRNA hairpin sequences as well as extra information including species of origin [19]. Additionally, the data introduced in version 22.1 was used as a testing set. We retrieved all 38589 (version 21: 28646) stem-loop microRNA samples from 271 species (version 21: 223) in the entire collection (including version 22.1) [19]. There are even more samples in miRBase version 22 and 21, although the precursor of Aves, Brassicaceae, Cercopithecidae, Embryophyta, Fabaceae, Hexapoda, Hominidae, HomoSapiens, Laurasiatheria, Malvaceae, Monocotyledons, Nematoda, Pisces, Platyhelminthes, Rodentia, Viruses were used to analyze classification of miRNA-Species in this study after cleaning procedure.

#### 3.1.2 Dataset of miRNA-Disease

The dataset of relationships between human miRNAs and diseases was obtained from HMDD 3.2 [23]. The entire collection of miRNA-disease association data (version 2019.01) was downloaded, which contains 35547 experimentally proven human miRNA-disease relationships for 222 disorders and 1206 miRNAs. We describe our data cleaning and preparation approach in following section, where we deleted duplicated miRNAs in a disease and chose diseases with a large number of linked miRNAs. The diseases of Carcinoma-Hepatocellular, Breast Neoplasms, Colorectal Carcinoma, Gastric Neoplasms, Lung Neoplasms, Prostate Neoplasms, Carcinoma,

Lung-Non-Small-Cell, Melanoma, Ovarian Neoplasms, Glioma, Glioblastoma, Pancreatic Neoplasms, Carcinoma, Breast, Carcinoma, Renal Cell, Osteosarcoma, Heart Failure with more than 150 miRNAs were selected for miRNA-disease association in this thesis after cleaning process.

The miRNA sequences in this thesis were obtained from miRbase, and the disease-miRNA association data was obtained from the HMDD (Human miRNA Disease Database) 3.2 website [23]. MiRbase release 22.1 [5] provided datasets for matching the sequences of miRNAs with linked diseases. This sequence database contains entries in FASTA format that represent the predicted hairpin region of a miRNA transcript. The dataset was gradually classified according with taxonomic tree, with the aim of getting one clade of the taxonomic tree for each clade of the information. Then, during in the cleaning procedure, the human miRNAs were retrieved from the grouped dataset, as detailed in the following section.

### **3.2 Data Cleaning**

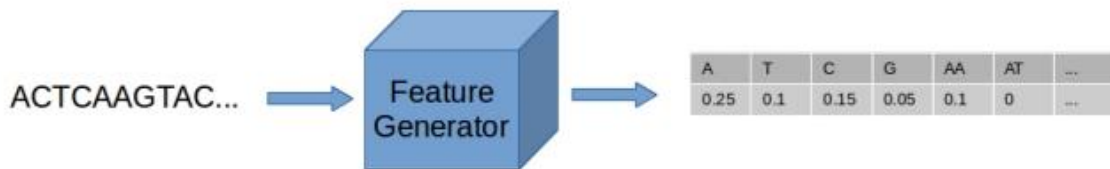
Because there were too many similar sequences and duplicates in the data, they were deleted from the dataset to avoid biasing the classifier. Every clade was cleaned separately for miRNA-Species data. The actual cleaning was done with the USEARCH tool, which clusters the data using k-medoid clustering, with two sequences belonging to the same cluster if they have a minimal similarity. The similarity criterion was set at 0.9, with 1 being totally equal. The medoids of the clustering result make up the cleaned dataset. The following was the USEARCH command that was used to clean: [input filename] usearch11.0.667 i86linux32 -cluster fast .txt] -centroids 0.9 -id [output file.fa].

To cluster the sequences based on their similarities, the USEARCH UCLUST method were employed. UCLUST groups together miRNA sequences that are related. The comparable sequences were clustered and replaced with their cluster medoids, which are different sequences used in the new dataset, after running the UCLUST algorithm.



### 3.3 Features

MicroRNA sequences can be represented in a variety of ways for use in machine learning algorithms [20][25][26]. The k-mer format, which is used in sequence analysis and computational genomics, is a well-known example of these representations. The collection of features in the k-mer representation is formed by adding all sequences' subsequences of required length. The 2-mer representation of nucleotides from the letters A, U, C, G, for example, yields 16 distinct characteristics such as AA, AU, GU, and so on. Similarly, nucleotide 3-mer representations create 64 distinct properties, such as AAA, AAC, AUC, and so on. We employed a mix of 1, 2, and 3-mers as features in this investigation, resulting in a total of  $4+16+64=84$  distinct characteristics for each miRNA.



**Figure 3.1** An example of transformation from miRNA sequences to k-mer frequency feature vectors

### 3.4 Sampling Methodology

Sampling methods are used to balance the size of two classes; they were required in this study because classification algorithms were trained on each class and owing to hierarchical clustering and the availability of miRNAs, each class may be of a wildly varying size. This may cause the classifier to favor the larger class. In background chapter can be seen more details. To balance the class sizes, the k-medoid sampling methods, and SMOTE sampling, which were discussed in the background chapter, were used in this study.

### **3.5 Choice of the Classifier**

The best classifier needs to choose in order to make hierarchical grouping of miRNA-Disease and miRNA-Species. In this study. Logistic Regression (LR), k-Nearest Neighbors (kNN), Naive Bayes (NB), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Multilayer Perceptron (MLP) classification models were compared to choose the best classifier.

The performance of each classifier is evaluated using 5-fold cross-validation [54]. This approach divides the data into ten disjoint subgroups at random. The model is then tested on one subset at a time, with the remaining sets being utilized to create the model.

The accuracy, precision, recall, F1 score, and log loss methods can be used to evaluate the classification performance. The total proportion of properly categorized observations is measured by accuracy, while the proportion of correctly assigned observations for each class is measured by recall. Precision is a performance metrics that indicates success in a positively predicted state, when F1 score is harmonic mean of recall and precision. In addition, log loss takes into account the uncertainty of the estimate based on how much it varies from the true value. This gives a more nuanced perspective on the performance of our model. Instead of another evaluation metrics, a lower log loss value represents better model prediction.

### **3.6 Postprocessing**

After choosing the best classification algorithm, the findings matrix of best classifier is transposed to provide a probability score for each species or diseases, which is then summarized in a membership probability vector (MP-vector). Each dimension of an MP-vector represents a species, with the values indicating the likelihood of belonging to that species.

**Table 3.1 A Part of MP-Vector**

	mirID 1	mirID 2	mirID 3	mirID 4	mirID 5	mirID 6	.....
<b>Specie/Disease 1</b>	0.093952	0.038179	0.081006	0.056612	0.091989	0.043164	.....
<b>Specie/Disease 2</b>	0.004146	0.060793	0.024256	0.065981	0.004187	0.033802	.....
<b>Specie/Disease 3</b>	0.186842	0.049447	0.117637	0.063065	0.204288	0.062946	.....
<b>Specie/Disease 4</b>	0.186842	0.049447	0.117637	0.063065	0.204288	0.062946	.....
<b>Specie/Disease 5</b>	0.176519	0.053156	0.132278	0.068295	0.204886	0.074003	.....
<b>Specie/Disease 6</b>	0.038691	0.027963	0.054718	0.053414	0.048702	0.029257	.....
<b>Specie/Disease 7</b>	0.186842	0.049447	0.117637	0.063065	0.204288	0.062946	.....
.....	.....	.....	.....	.....	.....	.....	.....

MP-vector is created for hierarchical clustering. Each class of data trained in the multi-classification algorithms represent rows in the MP-vector. Since the MP-vector is a probability vector, the sum of the values in each column is 1. The number of miRNAs in testing data is equal to the number of columns in MP-vector, while the number of classes is equal to the number of rows.

### 3.7 Choice of the Hierarchical Clustering

Hierarchical clustering is directly related to choosing the appropriate algorithm to make a good clustering for the dataset. Distance metrics and clustering methods generally try to understand the clustering algorithm that works very fast and efficiently in order to define the clustering design suitable for the structure of the dataset.

Distance metrics measures how similar two data points are. In most cases, all attributes of the data points contribute equally to the calculation of the closeness measure. No feature of the data points is dominant over the others. There are some distance metrics such as Euclidean, Canberra, Manhattan, Minkowski, Spearman, Pearson and Kendal distance metrics, although the Euclidean, Canberra, Manhattan and Minkowski distance metrics are used in this study.

Clustering method is necessary to define the clustering criterion, which can be expressed with a fixed function or other kinds of rules. It should be processed taking into account all the cluster types expected to occur in the dataset. Thus, it helps to determine the best clustering criterion that provides the correct division into the data set. There are some hierarchical clustering methods such as single linkage (Nearest Point algorithm), complete linkage (Farthest Point algorithm), average linkage (UPGMA algorithm), ward (Variance Minimization algorithm), centroid (UPGMC algorithm) and weighted (WPGMA) hierarchical clustering methods, although the single linkage, complete linkage, average linkage, ward and centroid clustering algorithms are used in this study.

When the studies are examined, it has been tried to determine the most accurate clustering method that can be applied to 45 different ceramic pieces collected from 3 different accident areas. As a result of the analysis, it was observed that the Cophenetic correlation coefficient achieved the highest value in the mean connection method [46]. In a different study, it was tried to determine the most accurate clustering method with the same method for 1560 accidents in 26 districts. It was observed that the cophenetic correlation coefficient achieved the highest value in the mean connection method [48]. In a different study, it was tried to determine the most accurate clustering method for 211 security design patterns. It was observed that the cophenetic correlation coefficient achieved the highest value in the mean connection method [51]. In the study, which was examined from different angles in the literature, different data sets were created according to the number of variables and the number of observations. For this ger data set, it was tried to find the best clustering method with the Cophenetic correlation coefficient. In all data sets, it was observed that the highest values were obtained in the mean connection method of the Cophenetic correlation coefficient [53].

As explained above, the Cophenetic correlation coefficient, which has proven its success in studies, will be the most appropriate coefficient choice to evaluate the compatibility between raw data distances and the distance measures used.

# Chapter 4

## Results and Discussions

### 4.1 Multi Class Classification

After cleaning the miRNA data obtained from bioinformatic databases in miRbase and HMDD with the UCLUST clustering algorithm using USEARCH, the cleaned data is converted into numerical data while it is a string with k-mer representation. In addition,

Logistic Regression (LR), k-Nearest Neighbors (kNN), Naive Bayes (NB), Linear Discriminant Analysis (LDA), Random Forest (RF), Decision Tree (DT) and Multilayer Perceptron (MLP) classification methods were applied cleaned original data and over-sampling minority class with SMOTE. Over sampling methods were applied to just minority classes of training dataset.

The parameters were tuned one by one on the data for all models without using any searching framework and cross validation, since trying the whole combination would require serious computing and time cost. Tuning parameters were implemented on the cleaned data whose features were converted numerical.

#### 4.1.1 miRNA-Disease Association

All classification methods which are selected was implemented in miRNA-Disease data. 'Breast Neoplasms', 'Carcinoma, Breast', 'Carcinoma, Hepatocellular', 'Carcinoma, Lung, Non-Small-Cell', 'Carcinoma, Renal Cell', 'Colorectal Carcinoma', 'Gastric Neoplasms', 'Glioblastoma', 'Glioma', 'Heart Failure', 'Lung Neoplasms', 'Melanoma', 'Osteosarcoma', 'Ovarian Neoplasms', 'Pancreatic Neoplasms' and 'Prostate

Neoplasms' were extracted because of having more than 150 miRNAs after cleaning part.

By adding all diseases, 3554 miRNAs were obtained, and all models were implemented with 5-fold cross-validation method on this data. Whole of classification models with tuned parameters were trained with class labels representing 16 diseases and 84 features obtained by mapping with k-mer representation.

**Table 4.2 LogLoss value of multi classification algorithms trained on cleaned miRNA-Disease data**

	CV 1	CV 2	CV 3	CV 4	CV 5	Average of CV
<b>LR</b>	2.7215	2.7299	2.7591	2.7489	2.7436	2.74060
<b>kNN</b>	2.7283	2.7331	2.7627	2.7662	2.7547	2.74900
<b>NB</b>	2.7792	2.7699	2.8173	2.8388	2.8217	2.80538
<b>LDA</b>	3.0097	3.0242	3.0389	3.0371	3.0337	3.02872
<b>RF</b>	2.7214	2.7301	2.7592	2.7490	2.7434	2.74062
<b>DT</b>	2.7215	2.7302	2.7592	2.7490	2.7434	2.74066
<b>MLP</b>	2.7267	2.7347	2.7578	2.7489	2.7432	2.74226

As can be seen in the Table 4.1, LR, kNN, RF, DT and MLP have taken very close values, and since low LogLoss value means better classification model, LR has become the first algorithm for the original data that has been cleaned with a 2.74060 LogLoss value.

MiRNA-disease data having 3554 miRNAs, which was created by combining all diseases, was increased minority classes by SMOTE method over the training set for each fold in cross validation. The number of rows of the training set on each fold has changed and the classification models have been trained on it.

**Table 4.3 LogLoss value of multi classification algorithms trained on over-sampled miRNA-Disease data**

	CV 1	CV 2	CV 3	CV 4	CV 5	Average of CV
<b>LR</b>	2.7726	2.7726	2.7726	2.7726	2.7726	2.77260
<b>kNN</b>	2.8271	2.8047	2.8050	2.8167	2.8325	2.81720
<b>NB</b>	2.9479	2.9274	2.9559	2.9605	2.9751	2.95336
<b>LDA</b>	3.1658	3.1819	3.1868	3.1795	3.1884	3.18048
<b>RF</b>	2.7727	2.7725	2.7724	2.7729	2.7725	2.77250
<b>DT</b>	2.7726	2.7726	2.7726	2.7726	2.7726	2.77260
<b>MLP</b>	2.7782	2.7742	2.7752	2.7738	2.7727	2.77482

LR, RF, DT and MLP have taken so close values, as be shown in the Table 4.2. RF has become the first algorithm for the over-sampled data with a 2.77250 LogLoss value.

Since classification algorithms trained on non-over-sampled data give better results, LR trained on original data is used for hierarchical clustering.

#### **4.1.2 miRNA-Species Association**

MiRNA-Species data is analyzed by all classification methods which are selected. 'Aves', 'Brassicaceae', 'Cercopithecidae', 'Embryophyta', 'Fabaceae', 'Hexapoda', 'Hominidae', 'HomoSapiens', 'Laurasiatheria', 'Malvaceae', 'Monocotyledons', 'Nematoda', 'Pisces', 'Platyhelminthes', 'Rodentia', 'Viruses' were extracted after cleaning and extracting features part.

14195 miRNAs were gained by adding all species and were trained with 5-fold cross-validation method using same classification model on this data. All classification models with tuned parameters were trained with class labels representing 16 diseases and 84 features obtained by mapping with k-mer representation.

**Table 4.4 LogLoss value of multi classification algorithms trained on cleaned miRNA-Species data**

	CV 1	CV 2	CV 3	CV 4	CV 5	Average of CV
<b>LR</b>	2.1143	2.1224	2.1497	2.1326	2.1689	2.13758
<b>kNN</b>	2.2214	2.221	2.2226	2.2121	2.2494	2.2253
<b>NB</b>	2.7132	2.6963	2.7210	2.6650	2.7321	2.70552
<b>LDA</b>	2.1745	2.1822	2.2399	2.1973	2.2348	2.20574
<b>RF</b>	1.9429	1.9523	1.9417	1.9658	1.9730	1.95514
<b>DT</b>	2.3544	2.3325	2.3205	2.3510	2.4009	2.35186
<b>MLP</b>	1.9769	1.9890	2.0075	2.0300	2.0393	2.00854

As can be seen in the Table 4.3, RF came first with the best LogLoss value compared to other methods for the original data that has been cleaned. RF take the average of LogLoss value equals to 1.95514 for 5-fold cross validation.

MiRNA-Species data having 14195 miRNAs, which was created by combining all species, was increased minority classes by SMOTE method over the training set for each fold in cross validation. The number of rows of the training set on each fold has changed and the classification models have been trained on it.

**Table 4.5 LogLoss value of multi classification algorithms trained on over-sampled miRNA-Species data**

	CV 1	CV 2	CV 3	CV 4	CV 5	Average of CV
<b>LR</b>	2.2960	2.3004	2.3376	2.3179	2.3528	2.32094
<b>kNN</b>	2.6406	2.6079	2.6045	2.5909	2.6507	2.61892
<b>NB</b>	3.2386	3.2198	3.2407	3.186	3.2168	3.22038
<b>LDA</b>	2.4234	2.4455	2.5267	2.4517	2.5004	2.46954
<b>RF</b>	2.0569	2.0679	2.0517	2.0797	2.0849	2.06822
<b>DT</b>	2.5079	2.5397	2.4659	2.5074	2.5170	2.50758
<b>MLP</b>	2.1399	2.1608	2.1561	2.2235	2.2118	2.17842



RF again came first with the best LogLoss value compared to other methods for the over-sampled data with a 2.06822 LogLoss value.

Since multi classification algorithms trained on non-over-sampled data give better results, RF trained on original data is used for hierarchical clustering.

## 4.2 Hierarchical Clustering

Following the selection of the best classification technique, the best classifier's results matrix is transposed to provide a probability score for each species or illness, which is then summarized in a MP-vector. After the MP-vector is created, the data is ready for hierarchical clustering.

All diseases and species were calculated all pairwise distance using Euclidean, Canberra, Manhattan and Minkowski distance metrics and Single, Average, Complete, Ward and Centroid was used as hierarchical clustering methods using same distance metrics. Then, Cophenetic distance were calculated for linkage which created from hierarchical clustering methods. Finally, the Pearson Correlation was calculated between pairwise distance and cophenetic distance.

### 4.1.1 miRNA-Disease Association

Logistic Regression (LR) gave the best LogLoss value after training all classification algorithms for miRNA-disease association. Due to insufficient miRNA-disease data, the test set size was set to 0.2 to increase the number of samples in the test set.

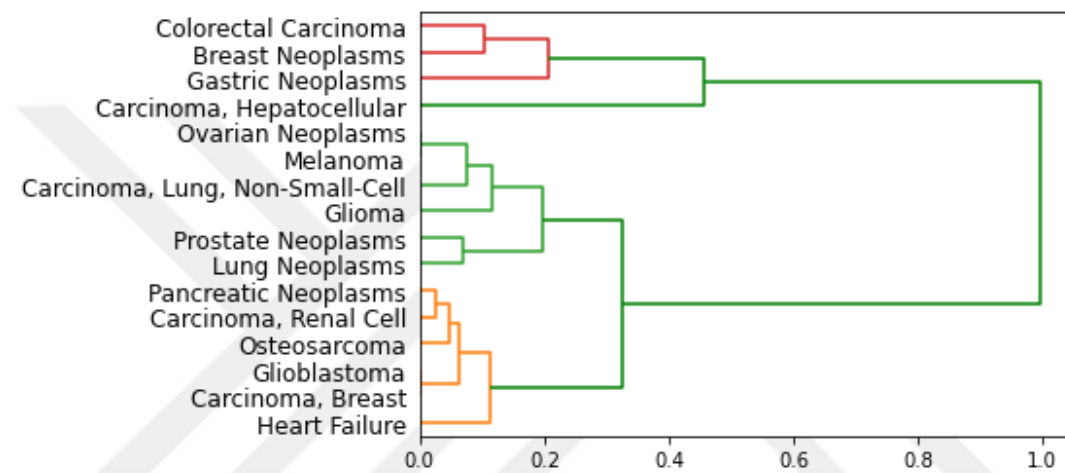
**Table 4.6 Comparison of clustering methods and distance metrics in the miRNA-Disease data**

		DISTANCE METRICS			
		Euclidean	Canberra	Manhattan	Minkowski
<b>CLUSTERING METHODS</b>	<b>Single</b>	0.87848	0.86804	0.87848	0.87848
	<b>Average</b>	0.89340	0.87970	0.89340	0.89340
	<b>Complete</b>	0.89123	0.87655	0.89123	0.89123
	<b>Ward</b>	0.88912	-	-	-

	<b>Centroid</b>	0.89340	-	-	-
--	-----------------	---------	---	---	---

As can be seen in the Table 4.5, Average hierarchical clustering methods with Euclidean, Manhattan and Minkowski distance metrics gave the highest result and also Centroid hierarchical clustering methods with Euclidean distance metrics gave the best result.

Diseases-Disease association calculated with clustering methods was visualized with a dendrogram graph is presented Figure 4.1.



**Figure 4.1 Dendrogram graph with Centroid clustering method, Euclidean distance metrics in the miRNA-Disease data**

When the Dendrogram graph of the case where the clustering method is Centroid and the distance metric is Euclidean, it is seen that the diseases are divided into 4 clusters with a distance value of 0.3 units. When these clusters are examined, Colorectal Carcinoma, Breast Neoplasms and Gastric Neoplasms are in a cluster together, Carcinoma Hepatocellular are in a cluster alone, Ovarian Neoplasms, melanoma, Lung Carcinoma, Glioma, Prostate Neoplasms and Lung Neoplasms are in the other cluster, the remaining other diseases are in the final cluster.

After the best classification algorithms having was implemented, the correlation value was calculated between diseases each other. The correlation graph between diseases with each other is presented in Figure 4.2.

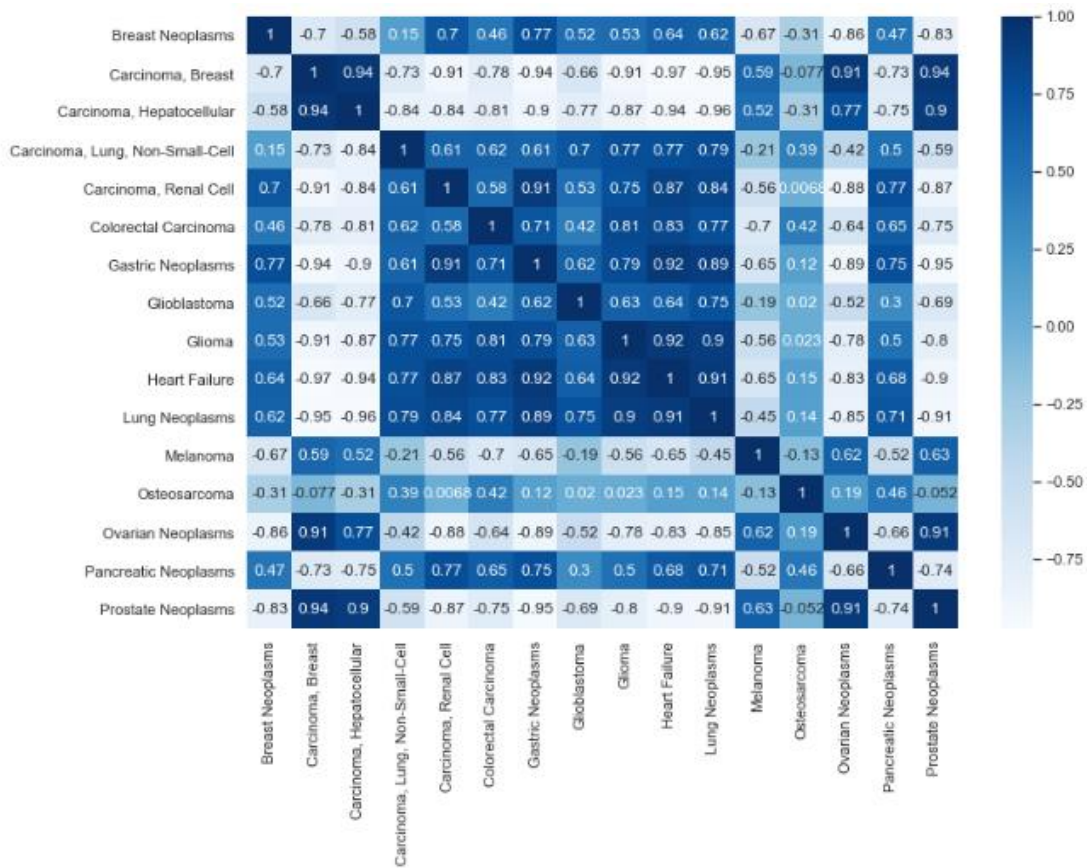


Figure 4.2 Correlation graph between diseases

#### 4.1.2 miRNA-Species Association

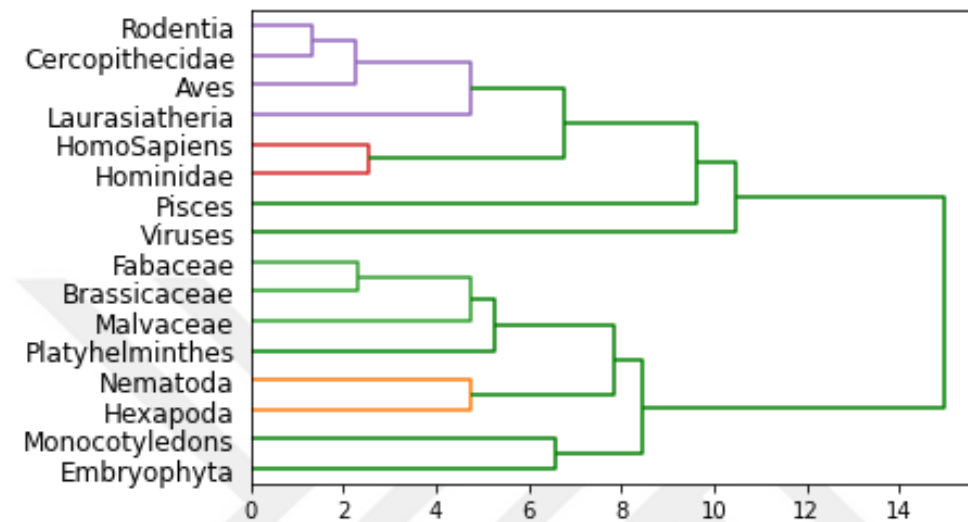
Random Forest (RF) has the lowest LogLoss value after training all classification algorithms for miRNA-Species association. Since there is not as little sample as in miRNA-Disease, the test set size was set to 0.1 in the test set for miRNA-Species data.

Table 4.7 Comparison of clustering methods and distance metrics on the miRNA-Species data

		DISTANCE METRICS			
		Euclidean	Canberra	Manhattan	Minkowski
CLUSTERING METHODS	Single	0.89153	0.97950	0.89569	0.89153
	Average	0.92728	0.98355	0.92044	0.92728
	Complete	0.92554	0.97984	0.91961	0.92554
	Ward	0.91126	-	-	-
	Centroid	0.9243	-	-	-

Pearson correlation in the miRNA-disease data set; It has been observed that Average clustering method and Canberra distance metric gives the highest result in the Table 4.6.

Species-Species association calculated with clustering methods was visualized with a dendrogram graph is presented Figure 4.3.



**Figure 4.3 Dendrogram graph with Average clustering method, Canberra distance metrics in the miRNA-Disease data**

When the Dendrogram graph of the case where the clustering method is Average and the distance metric is Canberra, it is seen that the diseases are divided into 8 clusters with a distance value of 6 units. When these clusters are examined, Rodentia, Cercopithecidae, Aves and Laurasiatheria are in a cluster together, HomoSapiens and Hominidae are in a cluster together, Pisces and Viruses are in a cluster separate and alone, Fabaceae, Brassicaceae, Malvaceae and Platyhelminthes are in a cluster together, Nematoda and Hexapoda are in the other cluster, Monocotyledons and Embryophyta are in a cluster separate and alone.

The correlation value was calculated between species each other, after the best classification algorithms was applied for miRNA-Species data. The correlation graph between species with each other is presented in Figure 4.4.

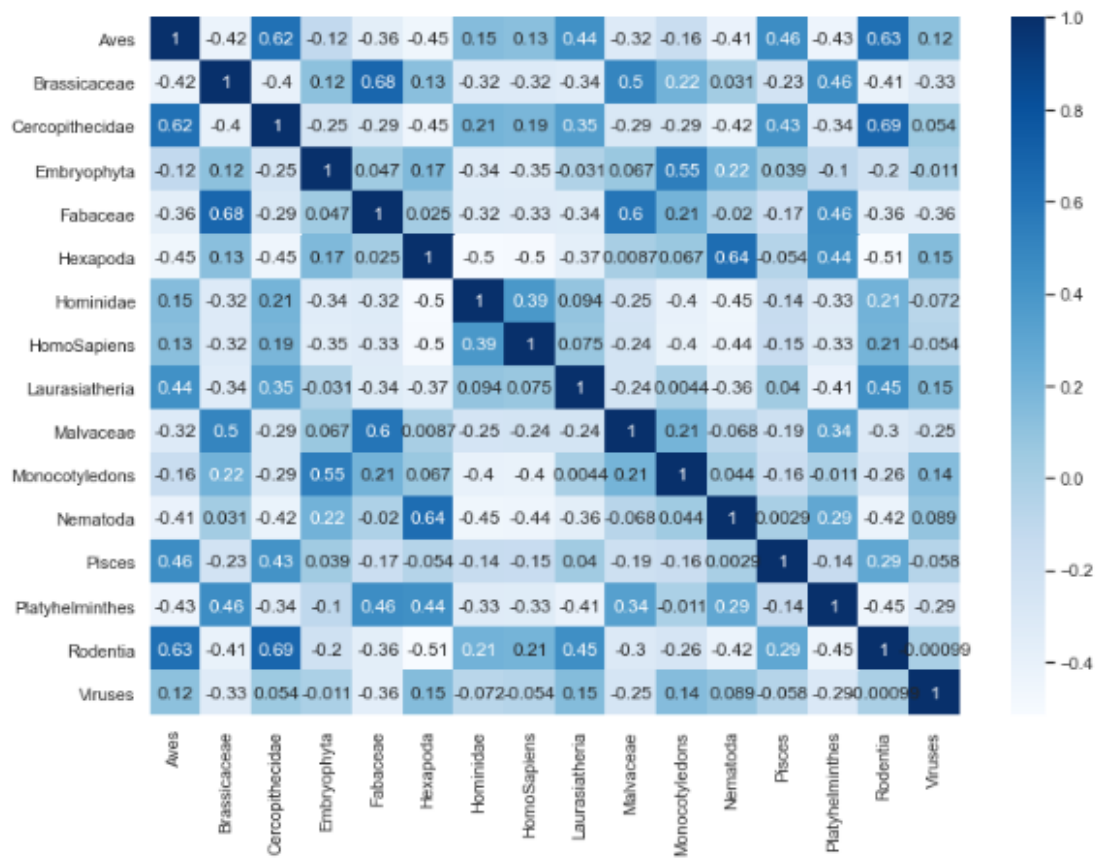


Figure 4.4 Correlation graph between species

# Chapter 5

## Conclusions and Future Prospects

### 5.1 Conclusions

Firstly, the UCLUST clustering algorithm was used to clean up similar miRNAs obtained from the miRBase database. Then, miRNA with proven disease relationships in the HMDD database was extracted in order to analyze the miRNA-disease associations.

In order to analyze both the miRNA-Disease associations and the miRNA-Species associations with a learning algorithm, miRNA sequences which consist of A, C, T, G letters had to be converted to numerical values. These miRNA sequences were converted into numerical data by k-mer frequency representation.

A problem that had to be faced was that the diseases and species were of different size, that is why over-sampling methods using SMOTE that deal with this problem. Since this method did not benefit the LogLoss score in the classification algorithms, SMOTE was not used and continued with the data that was non over-sampling.

In this study, 7 different classification algorithms were used. The classification method giving the best LogLoss value for both miRNA-Disease and miRNA-Species association was chosen, and the results were transposed to obtain a membership probability vector (MP-vector).

Finally, different hierarchical clustering methods with different distance metrics were applied on the MP-vector, and a dendrogram graph was created for miRNA-Disease association and miRNA-Species association by choosing the method and distance criterion that gave the highest correlation result.

One aim of this thesis was to design a new perspective that uses classification and clustering methods to classify miRNAs' associations with diseases and species. Another

goal of this project was to learn more about the evolution of miRNA and its differences across the phylogenetic tree in the species.

Classification and clustering methods can handle big data and are not limited to traditional rules for classifying miRNA as they learn classification rules automatically. This makes it possible to use the approach introduced in this thesis for automatic categorization of miRNA. Especially in the miRNA-Species approach, this study classifies a miRNA by its type of origin and if it is taxonomically related to the species presumed to originate, there may be a contamination and needs further examination.

## **5.2 Societal Impact and Contribution to Global Sustainability**

With the understanding of the role of miRNAs in disease development and formation and the improvement of molecular methods, many researches have been made and continue to be done in order to identify miRNAs specific to disease types and to develop new and effective treatment methods

By using the miRNAs involved in diseases and the sequence information of these miRNAs, the similarities of diseases and species with each other were determined by the classification and clustering methods developed throughout this thesis. This information can be used in drug repositioning studies, shedding light on the mechanisms of disease formation. In addition, miRNAs have been an important research focus due to their many different functions.

The discovery of a miRNA does not directly imply a specific function. Therefore, further analysis should be done. Some of these analyzes include identifying the source gene and its targets, as well as distinguishing real from pseudo miRNA. Although advancing bioinformatics technologies and next-generation sequencing make these miRNA identification tasks very easy to accomplish, they are still a complex and difficult task. The possible contaminated and pseudo miRNA problem in databases such as miRBase and the lack of a suitable method to find suspicious miRNAs have been an important motivation for this thesis.

It is thought that suspicious miRNAs originating from contamination or pseudo miRNA can be identified by classification and clustering methods. This will provide a

new method for ensuring the quality of miRNA databases such as miRBase. It can also help to learn more about miRNA evolution and differences across the phylogenetic tree. Basically, it categorizes miRNAs according to their source type so that miRNAs classified with another type than the one considered will be identified as likely suspect miRNA. In other words, if the predicted and target species are not related by a maximum distance, the miRNA in question will need to be examined further.

## **5.3 Future Prospects**

Despite the fact that the outcomes of this study are quite acceptable, there are certain aspects that may be improved. First of all, different mapping methods can be used instead of the k-mer frequency representation used for feature extraction. In addition, 7 different classification algorithms were used in this study. However, it has been shown that in many cases a combination of different classifiers can lead to better results, as each has its own advantages. Likewise, 5 different hierarchical clustering methods and 4 different distance metrics were used. Apart from these, many linkage methods and distance metrics can be tried and new dendrogram graphs can be created. In addition, limited species and diseases were used in this study due to computational costs. More comprehensive studies can be done by using all diseases and species found in databases. Thus, it is possible that especially miRNA-Species associations more advanced approaches, such as heuristics or taxonomic relationships between clades, will produce better results.



# BIBLIOGRAPHY

- [1] R. C. Lee, R. L. Feinbaum, and V. Ambros, “The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*,” *Cell*, vol. 75, no. 5, pp. 843–854, Dec. 1993, doi: 10.1016/0092-8674(93)90529-Y.
- [2] B. Wightman, I. Ha, and G. Ruvkun, “Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*,” *Cell*, vol. 75, no. 5, pp. 855–862, Dec. 1993, doi: 10.1016/0092-8674(93)90530-4.
- [3] B. J. Reinhart *et al.*, “The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*,” *Nature*, vol. 403, no. 6772, pp. 901–906, Feb. 2000, doi: 10.1038/35002607.
- [4] A. E. Pasquinelli *et al.*, “Conservation of the sequence and temporal expression of *let-7* heterochronic regulatory RNA,” *Nature*, vol. 408, no. 6808, pp. 86–89, Nov. 2000, doi: 10.1038/35040556.
- [5] P. B. Kwak, S. Iwasaki, and Y. Tomari, “The microRNA pathway and cancer,” *Cancer Sci.*, vol. 101, no. 11, pp. 2309–2315, Nov. 2010, doi: 10.1111/J.1349-7006.2010.01683.X.
- [6] Y. Lee *et al.*, “The nuclear RNase III Droscha initiates microRNA processing,” *Nat.* 2003 4256956, vol. 425, no. 6956, pp. 415–419, Sep. 2003, doi: 10.1038/nature01957.
- [7] E. Lund, S. Güttinger, A. Calado, J. E. Dahlberg, and U. Kutay, “Nuclear export of microRNA precursors,” *Science*, vol. 303, no. 5654, pp. 95–98, Jan. 2004, doi: 10.1126/SCIENCE.1090599.
- [8] H. Zhang, F. A. Kolb, V. Brondani, E. Billy, and W. Filipowicz, “Human Dicer preferentially cleaves dsRNAs at their termini without a requirement for ATP,” *EMBO J.*, vol. 21, no. 21, pp. 5875–5885, Nov. 2002, doi: 10.1093/EMBOJ/CDF582.
- [9] E. Bernstein, A. A. Caudy, S. M. Hammond, and G. J. Hannon, “Role for a bidentate ribonuclease in the initiation step of RNA interference,” *Nature*, vol. 409, no. 6818, pp. 363–366, Jan. 2001, doi: 10.1038/35053110.
- [10] R. I. Gregory, T. P. Chendrimada, N. Cooch, and R. Shiekhattar, “Human RISC

- couples microRNA biogenesis and posttranscriptional gene silencing,” *Cell*, vol. 123, no. 4, pp. 631–640, Nov. 2005, doi: 10.1016/J.CELL.2005.10.022.
- [11] B. Czech *et al.*, “Hierarchical rules for Argonaute loading in *Drosophila*,” *Mol. Cell*, vol. 36, no. 3, pp. 445–456, Nov. 2009, doi: 10.1016/J.MOLCEL.2009.09.028.
- [12] S. Yang *et al.*, “Widespread regulatory activity of vertebrate microRNA\* species,” *RNA*, vol. 17, no. 2, pp. 312–326, Feb. 2011, doi: 10.1261/RNA.2537911.
- [13] D. P. Bartel, “MicroRNAs: genomics, biogenesis, mechanism, and function,” *Cell*, vol. 116, no. 2, pp. 281–297, Jan. 2004, doi: 10.1016/S0092-8674(04)00045-5.
- [14] C. A. Andorfer, B. M. Necela, E. A. Thompson, and E. A. Perez, “MicroRNA signatures: clinical biomarkers for the diagnosis and treatment of breast cancer,” *Trends Mol. Med.*, vol. 17, no. 6, pp. 313–319, Jun. 2011, doi: 10.1016/J.MOLMED.2011.01.006.
- [15] G. Hutvagner and P. D. Zamore, “A microRNA in a multiple-turnover RNAi enzyme complex,” *Science*, vol. 297, no. 5589, pp. 2056–2060, Sep. 2002, doi: 10.1126/SCIENCE.1073827.
- [16] W. Filipowicz, S. N. Bhattacharyya, and N. Sonenberg, “Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?,” *Nat. Rev. Genet.*, vol. 9, no. 2, pp. 102–114, Feb. 2008, doi: 10.1038/NRG2290.
- [17] V. Ambros *et al.*, “A uniform system for microRNA annotation,” *RNA*, vol. 9, no. 3, pp. 277–279, Mar. 2003, doi: 10.1261/RNA.2183803.
- [18] S. Griffiths-Jones, “The microRNA Registry,” *Nucleic Acids Res.*, vol. 32, no. Database issue, Jan. 2004, doi: 10.1093/NAR/GKH023.
- [19] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright, “miRBase: microRNA sequences, targets and gene nomenclature,” *Nucleic Acids Res.*, vol. 34, no. Database issue, 2006, doi: 10.1093/NAR/GKJ112.
- [20] “miRBase.” <https://www.mirbase.org/index.shtml> (accessed Nov. 18, 2021).
- [21] Y. Li *et al.*, “HMDD v2.0: a database for experimentally supported human microRNA and disease associations,” *Nucleic Acids Res.*, vol. 42, no. Database issue, Jan. 2014, doi: 10.1093/NAR/GKT1023.
- [22] “HMDD v3.2.” <https://www.cuilab.cn/hmdd> (accessed Nov. 18, 2021).
- [23] I. H. Sarker, “Machine Learning: Algorithms, Real-World Applications and

- Research Directions,” *SN Comput. Sci.*, vol. 2, no. 3, May 2021, doi: 10.1007/S42979-021-00592-X.
- [24] “A simplified confusion matrix (Benos et al., 2021). | Download Scientific Diagram.” [https://www.researchgate.net/figure/A-simplified-confusion-matrix-Benos-et-al-2021\\_fig4\\_358823151](https://www.researchgate.net/figure/A-simplified-confusion-matrix-Benos-et-al-2021_fig4_358823151) (accessed Apr. 28, 2022).
- [25] J. Brownlee, “Master Machine Learning Algorithms Discover How They Work and Implement Them From Scratch i Master Machine Learning Algorithms,” 2016.
- [26] O. F.Y, A. J.E.T, A. O, H. J. O, O. O, and A. J, “Supervised Machine Learning Algorithms: Classification and Comparison,” *Int. J. Comput. Trends Technol.*, vol. 48, no. 3, pp. 128–138, Jun. 2017, doi: 10.14445/22312803/IJCTT-V48P126.
- [27] M. Usama *et al.*, “Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges,” *IEEE Access*, vol. 7, pp. 65579–65615, 2017.
- [28] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, 1990, doi: 10.1016/S0022-2836(05)80360-2.
- [29] R. C. Edgar, “Search and clustering orders of magnitude faster than BLAST,” *Bioinformatics*, vol. 26, no. 19, pp. 2460–2461, Aug. 2010, doi: 10.1093/BIOINFORMATICS/BTQ461.
- [30] M. Yousef, M. D. Saçar Demirci, W. Khalifa, and J. Allmer, “Feature Selection Has a Large Impact on One-Class Classification Accuracy for MicroRNAs in Plants,” *Adv. Bioinformatics*, vol. 2016, 2016, doi: 10.1155/2016/5670851.
- [31] S. Kurtz, A. Narechania, J. C. Stein, and D. Ware, “A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes,” *BMC Genomics*, vol. 9, p. 517, 2008, doi: 10.1186/1471-2164-9-517.
- [32] M. R. Longadge, M. Snehlata, S. Dongre, and D. Latesh Malik, “Class Imbalance Problem in Data Mining: Review,” *Int. J. Comput. Sci. Netw.*, vol. 2, no. 1, 2013.
- [33] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
- [34] “Pros and Cons of K-Nearest Neighbors - From The GENESIS.” <https://www.fromthegenesis.com/pros-and-cons-of-k-nearest-neighbors/> (accessed Apr. 29, 2022).

- [35] P. Kaviani and S. Dhotre, "Short Survey on Naive Bayes Algorithm," *Int. J. Adv. Res. Comput. Sci. Manag.*, 2017.
- [36] A. M. Mart Nez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, 2001.
- [37] C. Strobl, A. L. Boulesteix, and T. Augustin, "Unbiased split selection for classification trees based on the Gini Index," *Comput. Stat. Data Anal.*, vol. 52, no. 1, pp. 483–501, Sep. 2007, doi: 10.1016/J.CSDA.2006.12.030.
- [38] B. Taha Jijo and A. Mohsin Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.
- [39] J. M. Nazzal, J. M. Nazzal, I. M. El-etary, and S. A. Najim, "Multilayer Perceptron Neural Network (MLPs) For Analyzing the Properties of Jordan Oil Shale 1."
- [40] A. J. C. Witsil and J. B. Johnson, "Volcano video data characterized and classified using computer vision and machine learning algorithms," *Geosci. Front.*, vol. 11, no. 5, pp. 1789–1803, Sep. 2020, doi: 10.1016/J.GSF.2020.01.016.
- [41] S. Haykin, "Stochastic Machines and Their Approximates Rooted in Statistical Mechanics," *Neural Networks Learn. Mach.*, pp. 567–618, 2008.
- [42] K. L. Priddy and P. E. Keller, "Artificial Neural Networks: An Introduction," *Artif. Neural Networks An Introd.*, Sep. 2009, doi: 10.1117/3.633187.
- [43] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," *J. R. Stat. Soc. Ser. B*, vol. 36, no. 2, pp. 111–133, Jan. 1974, doi: 10.1111/J.2517-6161.1974.TB00994.X.
- [44] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview, II," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 7, no. 6, p. e1219, Nov. 2017, doi: 10.1002/WIDM.1219.
- [45] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, "Cluster analysis: Fifth edition," *Clust. Anal. Fifth Ed.*, pp. 1–330, Jan. 2011, doi: 10.1002/9780470977811.
- [46] P. R. Carvalho, C. S. Munita, and A. L. Lapolli, "Validity studies among hierarchical methods of cluster analysis using cophenetic correlation coefficient," *Brazilian J. Radiat. Sci.*, vol. 7, no. 2A, Feb. 2019, doi: 10.15392/BJRS.V7I2A.668.
- [47] M. S. Aldenderfer and R. K. Blashfield, "Cluster Analysis (Quantitative

- Applications in Social Sciences),” pp. 1–88, 1984.
- [48] S. Kumar and D. Toshniwal, “Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC),” *J. Big Data*, vol. 3, no. 1, pp. 1–11, Dec. 2016, doi: 10.1186/S40537-016-0046-3/FIGURES/4.
- [49] A. Ziviani, S. Fdida, J. F. De Rezende, and O. C. M. B. Duarte, “Toward a Measurement-Based Geographic Location Service,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3015, pp. 43–52, 2004, doi: 10.1007/978-3-540-24668-8\_5.
- [50] S. Kumar, “Performance Evaluation of Distance Metrics in the Clustering Algorithms | Scinapse,” 2014.
- [51] P. Ponde, S. Shirwaikar, and S. Gore, “Hierarchical cluster analysis on security design patterns,” *ACM Int. Conf. Proceeding Ser.*, vol. 12-13-August-2016, Aug. 2016, doi: 10.1145/2979779.2979871.
- [52] S.-S. Choi, S.-H. Cha, and C. C. Tappert, “A Survey of Binary Similarity and Distance Measures.”
- [53] S. Saraçlı, N. Doğan, and I. Doğan, “Comparison of hierarchical cluster analysis methods by cophenetic correlation,” *J. Inequalities Appl.*, vol. 2013, Apr. 2013, doi: 10.1186/1029-242X-2013-203.
- [54] R. Kohavi and R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” pp. 1137--1143, 1995.

# CURRICULUM VITAE

2007 – 2011 High School., Kayseri Science High School, Kayseri,  
TURKEY

2011 – 2017 B.Sc., Industrial Engineering, Sabanci University, Istanbul,  
TURKEY

2018 Certificate of Expertise, Big Data and Business Analytic, Istanbul  
Technical University, Istanbul, TURKEY

2019 – 2022 M.Sc., Electrical and Computer Engineering, Abdullah Gül  
University, Kayseri, TURKEY

## SELECTED PUBLICATIONS AND PRESENTATIONS

**C1)** M. Yousef, Y. Erbasi, B. Bakir-Gungor, Classification of microRNA Disease Association Based on k-mer Sequence Representation Only APBC2021, The 19th Asia Pacific Bioinformatics Conference (2021).