Gökhan GÖY

A Master's Thesis

AGU 2021

MACHINE LEARNING BASED INTEGRATION OF miRNA AND mRNA PROFILES COMBINED WITH FEATURE GROUPING AND RANKING

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
Gökhan GÖY
September 2021

MACHINE LEARNING BASED INTEGRATION OF miRNA AND mRNA PROFILES COMBINED WITH FEATURE GROUPING AND RANKING

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

Gökhan GÖY

September 2021

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.

Gökhan GÖY

REGULATORY COMPLIANCE

M.Sc. thesis titled **"MACHINE LEARNING BASED INTEGRATION OF miRNA AND mRNA PROFILES COMBINED WITH FEATURE GROUPING AND RANKING"** has been prepared in accordance with the Thesis Writing Guidelines of the Abdullah Gül University, Graduate School of Engineering & Science.

Prepared By	Advisor
Gökhan GÖY	Assistant Professor
	Burcu BAKIR GÜNGÖR

Head of the Electrical and Computer Engineering Program

Associate Professor Kutay I. İçöz

ACCEPTANCE AND APPROVAL

M. Sc. thesis titled **"MACHINE LEARNING BASED INTEGRATION OF miRNA AND mRNA PROFILES COMBINED WITH FEATURE GROUPING AND RANKING"** and prepared by Gökhan Göy has been accepted by the jury in the Electrical and Computer Engineering Graduate Program at Abdullah Gül University, Graduate School of Engineering & Science.

02 / 09 / 2021

JURY:

Assistant Professor Burcu BAKIR GÜNGÖR :.....

Assistant Professor Ufuk NALBANTOĞLU :.....

Associate Professor Mete ÇELİK :.....

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 ……..……. .

…….. /…….. / ……..

Graduate School Dean

Prof. Dr. Hakan USTA

ABSTRACT

MACHINE LEARNING BASED INTEGRATION OF miRNA AND mRNA PROFILES COMBINED WITH FEATURE GROUPING AND RANKING

Gökhan GÖY
MSc. in Electrical and Computer Engineering
Advisor: Assistant Professor Burcu BAKIR GÜNGÖR
September 2021

It is very important to understand the development and progression mechanisms of the diseases at the molecular level. Revealing the functional mechanisms that cause the disease not only contributes to the molecular diagnosis of the diseases, but also contributes to the development of the new treatment methods. Nowadays, due to the advances in technology, more molecular data can be obtained at cheaper costs, unlike in the past. Integrating these available data is essential to understand the molecular mechanisms of the diseases, especially the ones having complex formation and progression processes such as cancer.

In this thesis, to correctly classify cancer patients and cancer free patients, two different bioinformatics tools (miRcorrNet and miRMUTINet) that integrate mRNA and microRNA data (two types of -omic data at the molecular level) have been developed. For 11 cancer types, mRNA and miRNA expression profiles of the samples were downloaded from The Cancer Genome Atlas. These two data types were integrated using both the Pearson Correlation Coefficient and the Mutual Information metrics. In our experiments using 100-fold Monte Carlo Cross Validation, for both tools, 99% Area Under the Curve score have been obtained. The developed tools have also been tested using independent dataset. For biological validation purposes, for each cancer type, functional enrichment analysis is conducted on the identified list of significant miRNAs and genes. Additionally, for each cancer type, the identified mRNAs and miRNAs were subject to literature validation and the findings were noteworthy.

Keywords: Machine Learning, Classification, Grouping, miRNA, mRNA

ÖZET

ÖZELLİK GRUPLAMASI VE SIRALAMASI İLE BİRLİKTE miRNA VE mRNA EKSPRESYON PROFİLLERİNİN MAKİNE ÖĞRENİMİ TABANLI ENTEGRASYONU

Gökhan GÖY
Elektrik ve Bilgisayar Mühendisliği Anabilim Dalı Yüksek Lisans
Tez Yöneticisi: Dr. Öğr. Üyesi Burcu BAKIR GÜNGÖR
Eylül 2021

Hastalıkların oluşum ve gelişim mekanizmalarını moleküler seviyede anlamak çok önemlidir. Hastalığa yol açan fonksiyonel mekanizmaların açığa vurulması, yalnızca hastalıkların moleküler tanısına değil, aynı zamanda yeni tedavi yöntemlerinin geliştirilmesine de katkıda bulunur. Bugünlerde, teknolojideki ilerlemeler sayesinde moleküler veriler eski zamanların aksine daha ucuz fiyatlarla elde edilebilir. Bu erişime açık verilerin entegre edilmesi, özellikle kanser gibi kompleks oluşum ve ilerleme mekanizması olan hastalıkların moleküler mekanizmalarını anlamak için elzemdir.

Bu tezde, kanser hastalarını doğru sınıflandırmak için, mRNA ve mikroRNA verilerini (moleküler seviyede iki tip –omik veri) entegre eden miRcorrNet ve miRMUTINet adında iki adet araç geliştirildi. 11 kanser tipi için, örneklerin mRNA ve miRNA ekspresyon profilleri, The Cancer Genome Atlas'tan indirildi. İki veri tipi, hem Pearson Korelasyon Katsayısı, hem de Ortak Bilgi metrikleri kullanılarak entegre edildi. 100 katlı Monte Karlo Çapraz Doğrulama kullandığımız deneylerimizde, her iki araç için de 99% Area Under the Curve skoru elde ettik. Geliştirilen yöntemler bağımsız veri kümeleri ile de test edildi. Biyolojik doğrulama amacıyla, her kanser tipi için, önemli olduğu belirlenen miRNAlar ve genler listesi üzerinde, fonksiyonel zenginleştirilme analizi gerçekleştirildi. Ayrıca, her kanser tipi için, hastalıklar ile ilgili olduğu düşünülen mRNA ve miRNA'ler literatür validasyonuna tabi tutulmuş ve bulguların dikkate değer olduğu görülmüştür.

Anahtar Kelimeler: Makine Öğrenmesi, Sınıflandırma, Gruplandırma, miRNA, mRNA

Acknowledgements

I would like to thank my esteemed advisor, Burcu Bakir-Gungor, who enabled me to carry out this thesis and provided me with an important experience. I will always feel the happiness of working with you on a subject that can contribute to humanity.

I would like to give a special thanks to Prof. Dr. Malik Yousef. As a Co-Advisor, you gave me a great research experience. Thanks to you, I had the courage to go to unfamiliar lands. Now I am proud of being able to complete this journey. Working with you has taught me how to be a happier person in life: hardworking, passionate, focused. I know that I can't thank you enough, and I offer you my deepest respects.

I would like to give a special thanks to my dear colleague M. Furkan Akkoyunlu for his morphological advice and guidance during the writing of my thesis.

Of course, the most important thank you is to my very precious mother and my precious sister. For always being by my side and always supporting me no matter what happens in difficult times. I will be there for you.

May the force be with me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

WHO	World Health Organization
RNA	Ribo Nucleic Acid
miRNA	Micro RNA
mRNA	Messenger RNA
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast Invasive Carcinoma
KICH	Kidney Chromophobe
KIRC	Kidney Renal Clear Cell Carcinoma
KIRP	Kidney Renal Papillary Cell Carcinoma
LUAD	Lung Adenocarcinoma
LUSC	Lung Squamous Cell Carcinoma
PRAD	Prostate Adenocarcinoma
STAD	Stomach Adenocarcinoma
THCA	Thyroid Carcinoma
UCEC	Uterine Corpus Endometrial Carcinoma
RPM	Reads per Million
RPKM	Reads per Kilobase Million
MCC	Matthews Correlation Coefficient
MCCV	Monte Carlo Cross Validation
TCGA	The Cancer Genome Atlas

Chapter 1

Introduction

1.1 Motivation

The World Health Organization (WHO) estimates that cancer is the first or second leading cause of death in almost three-fourths (3/4) of the world countries [1]. It is estimated that there were more or less 19.3 million new cancer cases in 2020 and the number of deaths was roughly 10 million. The number of new cases and the number of deaths for 36 types of cancer is given in Figure 1.1. Cancer in 57 countries, including Turkey, is considered to be disease that causes most deaths among diseases. On the other hand, cancer is the second leading cause of death in 55 countries. In 23 countries, it is the third most common disease that causes death. In the light of all this information, cancer is thought to be the first among diseases that cause death in the world.

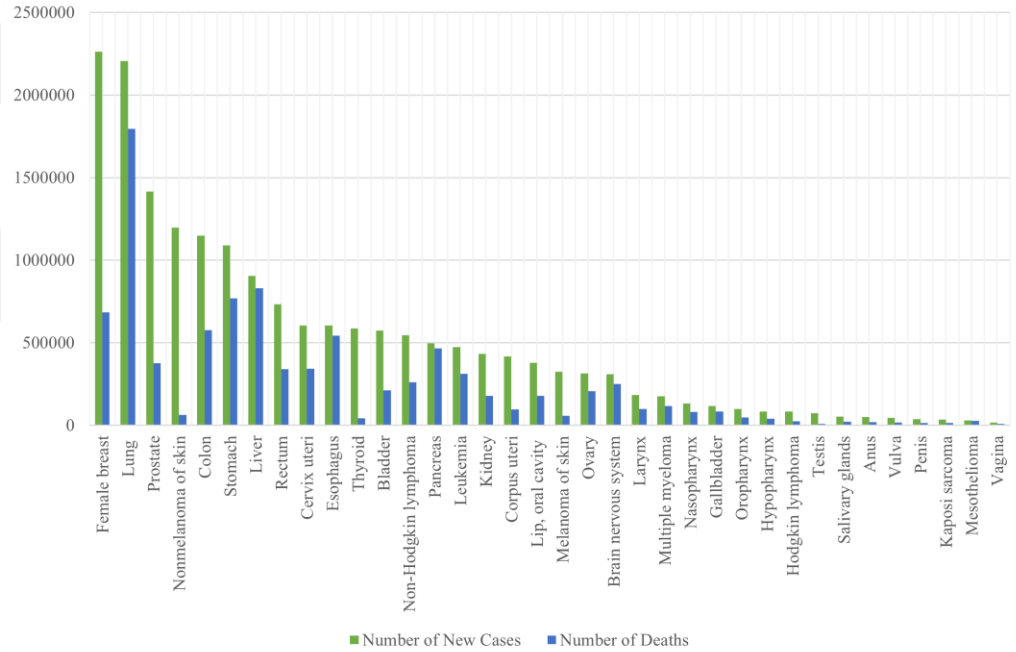

Figure 1.1 All Cancer Cases and Mortality Information

Technology is developing dramatically and rapidly. In this regard, the formation of the disease, its progression and spread of the disease, in short, molecular mechanisms of the diseases can be studied more easily today. Compared to the times when technology was relatively limited, much more data of the same patient can be produced using the new generation high-volume technologies with cheaper costs. Various paradigms have been developed in order to make it possible to understand the mechanisms of diseases at a more complex level, even if not for every disease, by using this vast amount of data. For example, single-omic data was used primarily to understand the mechanisms of diseases. However, with the increase of the available data over time, the use of multi-omics techniques has increased. It is also clear that after the use of multi-omics technologies has matured and reached state-of-the-art, the trend will evolve into personalized medicine.

Bio-validation of information obtained using data is essential in health industry. For this purpose, scientists carry out wet-lab operations. However, these operations have some shortcomings. First of all, the cost of these procedures is very high. In addition, the time required to complete these processes is often long. Additionally, qualified experts are required to perform these processes, which poses a huge problem for most middle and low-income countries. Therefore, biological control and approval of too much data is not feasible due to above-mentioned reasons. Thus, obtained vast amount of data should also be used very effectively. Based on this information, it is very important to generate smart solutions to reduce the potentially biologically important information that needs to be validated by using state-of-the-art machine learning methods.

Until today, within the scope of the smart solutions mentioned above, tools that integrate miRNA and mRNA using various programming languages have been developed. However, these tools have various general disadvantages. For example, these tools are not updated as often as they need to, so they fall behind today's new technologies. Another point is that the use of these tools is complex, and they require too many third-party applications i.e., libraries and their adjustments are hard to follow. It should not be forgotten that people working in these fields, that is, clinicians and biological experts, have encounter various difficulties in using these tools. A bioinformatics tool to be developed should be easy to use, and it should keep up with

today's technologies, and should be in a structure that can appeal to experts in different fields.

1.2 miRNA

In the most general terms, RNAs are divided into 2 classes as coding RNAs and non-coding RNAs. miRNAs are non-coding RNAs roughly 22-25 nucleotides in length[2] There are some features that distinguish miRNAs from other types of non-coding RNAs. For example, miRNAs are preserved in various living species, without much change in their sequences. This can be shown as evidence that miRNAs play critical roles[3]. Although miRNAs have a single stranded structure and do not have a coding function, they regulate various and critical biological processes. Examples of these biological processes are events such as apoptosis [4], cell proliferation [4], or regulation of biological signaling pathways [5]. In addition to all these, it is now a known fact that miRNAs are active in all stages, including the formation of the disease of many human diseases [6]. It is also known that miRNAs play important roles in tangled diseases such as cancer [7]. The difficulties in these types of diseases are molecular mechanisms, all stages of a disease and the relationships between all biological molecules. Also, it has been stated in many studies that miRNAs can also be an ideal biomarker for diagnosis of cancer and an important molecule to be able to develop target medicines to cure diseases [8,9].

1.3 mRNA

Another class of RNAs is messenger RNAs (mRNAs). This class of RNA is also known as coding RNA. From the point of central dogma of biology, the production of mRNAs is as follows. Genes are translated into pre-RNA by means of enzymes and pre-RNA composed of exons and introns. Exons are segments that will be used in protein production. Again, by means of enzymes, mRNAs are formed by combining exons and shedding introns [10]. Hence, mRNA contains the codes for the protein that needs to be produced. Then, the synthesis of the relevant protein is carried out using this information. In short, it has a very important place at the molecular level. The important thing about mRNAs is that the more mRNAs in a sample, the more gene expression there is. In other words, the production of mRNA is expressed as gene activity, i.e., gene expression. Undoubtedly, there are relationships at the molecular level between

mRNAs (genes) and miRNAs. All the aforementioned events are provided by these relationship mechanisms.

1.4 miRNA - mRNA Relationship

There are mRNAs and miRNAs in most of the events that take place in the human body. These two molecules work together by cooperating at the molecular level. They perform these operations by affecting each other. In this context, this interaction relationship can be expressed as follows in the most general terms. While a miRNA can affect more than one mRNA, it is similarly known that more than one miRNA can affect only one mRNA. This situation enables the relationship to evolve into a many to many (m-n) relationship [11]. Understanding these many-to-many relationships is thought to be the key to understanding tangled diseases' mechanism such as cancer mechanisms. The benefits of knowing that the relationships exist between miRNAs and mRNAs are not only important to understanding their mechanism. It also makes a great contribution to the production of medicines that can potentially treat complex diseases [12]. miRNAs integrate itself to its target mRNAs by binding to motifs ie. seed region in the 3' UTR region in order to regulate its expression [13,14]. In the literature, it has been determined that this seed region has critical importance in determining miRNA and its target [15]. This process is carried out at the post transcriptional level. In this way, miRNAs induce mRNA degradation by suppressing gene expression. miRNAs generally tend to have an inverse expression effect on mRNAs. In other words, gene expression is often negatively correlated with its target miRNA(s). Based on this information, it is stated that the most important regulating power on mRNA is miRNAs [16].

1.5 Aim and Design of the Study

In this thesis, it is aimed to integrate miRNA and mRNA data using machine learning methods and to explore important miRNAs and mRNAs and classify patient with cancer and patient without cancer people through this machine learning model. In addition to the individual effects of the genes and miRNAs used in the classification process, the effects of groups formed by miRNA and its target mRNA(s), in other words clusters, were also examined. With this way, it was aimed to determine statistically significant miRNAs, mRNAs and the groups of these for the current disease under

study. By using multi-omics data, it was aimed to provide a better understanding of the mechanisms of tangled diseases.

With this thesis, it has become possible to use the vast amount of publicly available data in the most effective and efficient way. By means to the acquired outputs, the number of molecules that need to be validated biologically has been tried to be minimized. In this way, it was tried to get rid of the disadvantages of wet-lab operations as much as possible. In addition, with the use of the developed machine learning model, it is aimed to identify molecules that are more important for the disease and contribute not only to diagnosis but also to treatment approaches. In addition to all these, it is aimed to produce a user-friendly bioinformatics tool that is compatible with the latest technology and is extremely easy to use.

In the thesis design phase, first of all, finding statistically important mRNAs and miRNAs was prioritized. Subsequently, mRNAs were grouped based on miRNAs and miRNA-mRNA clusters were obtained. Then, the success of these groups in separating two classes was measured with various metrics and scores belonging to each group were obtained and ranked. Finally, with this way, a robust machine learning model that can detect the most effective groups using the least gene and miRNA has been obtained. To prove the reliability of the obtained model, the literature validation method was adopted, and it was shown that the outputs were robust.

The rest of this thesis organized as follows. In chapter 2, studies that perform miRNA-mRNA integration have been mentioned. These studies have been examined to include all the different methods found in the literature. In addition, the tools used in the literature are also mentioned in order to compare the results. In Chapter 3, all used materials in this thesis have been explained. In addition, the solution approaches of 2 developed tools namely miRcorrNet [17] and miRMUTINet have been mentioned. In chapter 4, all obtained results are presented for both miRcorrNet and miRMUTINet, and all bioinformatics tools are compared against various performance metrics. In chapter 5 all the results have been discussed. In chapter 6, the aims of the thesis, the experiments carried out and the findings obtained are evaluated. In addition, the impact of the thesis on society was evaluated. Finally, the addresses of the repository containing all the experiments and all their materials are shared.

Chapter 2

Literature Review

In this section, the methodological techniques used for miRNA-mRNA integration in the literature will be discussed. In general terms, studies in the literature can be seen in Figure 2.1. Studies use different type of methodologies to integrate miRNA and mRNA. These methodologies can be divided into 6 different techniques. These techniques are correlation-based techniques, linear modeling-based techniques, Bayesian network-based techniques, Statistical method-based techniques, probability-based techniques, and Matrix Decomposition-based techniques, respectively. Each technique, and the publications which use these techniques will be discussed.

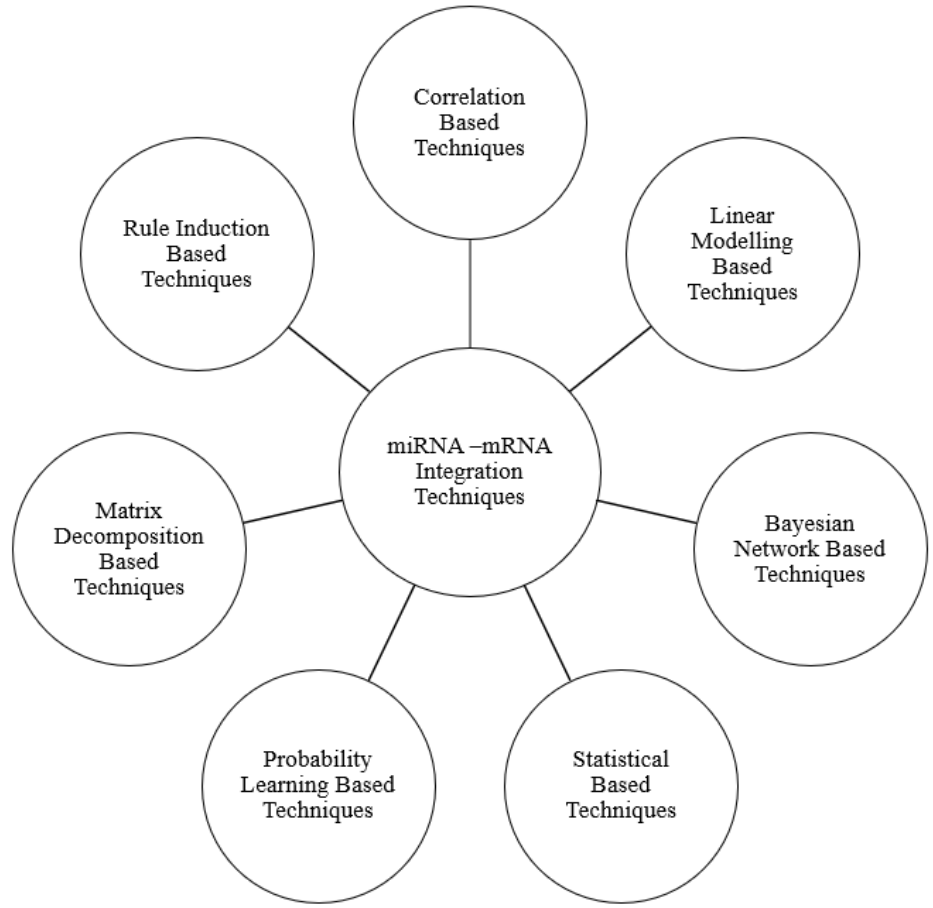

Figure 2.1 Utilized Techniques for miRNA-mRNA Integration

2.1 Correlation Based Techniques

Correlation-based techniques are non-complex and relatively easy to perform. Correlation metric reveals whether there is a relationship in other words dependence between two features within the data. The dependence is determined using correlation value. In this context, the range of correlation values can be specified as [-1, +1]. In general terms 3 values have interpretation. Value of -1 indicates that these two features have a firm negative correlation dependence. Value of 0 indicates that there is no dependence between these two features. And value of +1 indicates that they have a strong positive correlation dependence. In these methods, all miRNA-mRNA pairs are evaluated in a pairwise manner. It is aimed to determine the associated pairs by calculating the linear correlation coefficient values of each miRNA-mRNA. All available miRNA-mRNA pairs are not used to integration of miRNA and mRNA data. It is necessary to obtain pairs that are considered to be important for this integration. It has been shown in the literature that miRNAs affect mRNAs. This relationship has often been expressed in the literature as miRNAs are negatively correlated its target mRNA(s). However, in one study, it was claimed that miRNAs up-regulate mRNAs [18]. Because of this, researchers tend to use negatively correlated miRNA-mRNA pairs to integrate these two different data. Although correlation solutions are a bit more practical, they have a disadvantage. This disadvantage is that correlation value calculations are done in a pairwise manner. In other words, correlation-based solutions presume that one miRNA affects only one mRNA, which is not always the case as mentioned earlier.

Feng et al. conducted a study using the mRNA-miRNA integrated analysis method so as to understand the biological and molecular mechanism of cardiac dysfunction in rats with selenium deficiency [19]. Liu et al. sought to reveal the associations leading to the formation of curly toison using the mRNA-miRNA integrated analysis method [20]. Li et al. aimed to better understand the mechanisms leading to aggressive Lung Adenocarcinoma (LUAD), which has a low survival rate, by separating metastatic and non-metastatic LUAD patient groups [21]. The authors identified candidates for prognostic markers and potential targets for LUAD therapy.

Yao et al. identified potentially important several mRNA-miRNA pairs and signaling pathways for atrial aging to improve understanding of atrial aging at the molecular level [22]. Yang et al. conducted studies using mRNA-miRNA integrated analysis to increase understanding into the mechanisms of right sided and left sided Colon Adenocarcinoma [23].

2.2 Linear Modelling Based Techniques

The linear model technique is a method that arises due to a lack of correlation-based approaches. As will be remembered, it has been stated that there are many to many relationships between miRNAs and mRNAs. Because correlation-based systems focus on finding miRNA-miRNA pairs. So, those techniques do not evaluate many-to-many relationships. With this technique, in order to overcome the deficiency caused by this situation, the effects of miRNAs on a single mRNA are investigated by considering miRNAs as independent variables and mRNA as dependent variable.

Huang JC et al. endeavored to model mRNA expressions as linear combinations of miRNAs in their study [24]. However, they also used the Bayesian algorithm to discover hidden miRNA targets. Later they used a different distribution technique, integrating sequence information, to their previous study [25]. Stingo et al. also proposed a comparable solution approach in their study [26]. Differently, they did not consider the tissue effect and stated that miRNAs had a different promoter effect on each mRNA. In [27], the authors tried to find the mRNA modules that affect the functioning of miRNAs by performing miRNA - mRNA integration. They used both interaction, expression and sequence information for a regression-based solution to ensure this integration. The authors claimed that using this method, they found the modules in a much more robust and accurate way. In this study, the process begins with converting all the different data into matrix format as preprocessing step. Subsequently, the desired number of modules is specified. In this way, it is considered to determine the members of the modules. The coefficients required for the regression model were obtained with some presumes. Then, a function for optimization is generated for the integration of all the acquired information. According to the information the authors

gave, function can encounter with local minima problem. In order to solve this problem, various parameters were added to this function and then finally optimized modules were obtained. CoModule is another study which tries to find mRNA modules [28]. This study, which adopts a grouping-based method with the integration of different data, first groups miRNAs according to similar expression values. Then, final modules are obtained by adding associated mRNAs with the respective miRNA(s). Different data were used with lasso regression model to generate miRNA-mRNA pairs.

2.3 Bayesian Network Based Techniques

Bayesian network-based solutions are also available in the literature as another technique on this subject. In some studies researchers had used this technique for similar objectives and they got successful results [29]. This situation led to have opened up the use of this technique. Although the Bayesian network technique produces reliable results [26], it is not feasible to use it in large networks because this technique utilizes exhaustive searching algorithms in the background. Liu et al. used the Bayesian network technique, to make mRNA-miRNA integrated analysis [30]. In this study, the authors first tried to find differentially expressed miRNAs and mRNAs using their pValues. To be able to use this data, first they discretized the data as preprocessing. Later, they used the bayes network learning method to find relationships between miRNAs and mRNAs. Then, to ensure the robustness of the findings they used target information. Finally, they merged all the networks in order to acquire miRNA- mRNA network. In addition, they have used some conditions in the search field to compensate for the weakness of the Bayesian network technique.

2.4 Statistical Based Techniques

Another technique used for miRNA-mRNA integration is statistical methods. Events that take place at the molecular level do not happen by chance. On the contrary, all events that take place in a living thing's system, take place with a specific biological systematic. This situation triggered the tendency towards statistical based solutions in the field of bioinformatics. Because output of bioinformatic should be biologically relevant. These methods are based on some basic presumes and enable to acquire robust results. In this study [31], a grouping-based method was used to find miRNA-mRNA modules. In the study, firstly miRNA and mRNA clusters were found in a separate

manner. Afterwards, the relationships between these groups were analyzed statistically and the final modules were tried to be created by merge the groups that were considered to be statistically significant. An enrichment analysis was performed to create the groups, and then the existence of the relationships was tested linearly according to various conditions. In this study [32], the authors developed an approach to distinguish different tissues. Using this method, it does this by simply looking at different miRNA-mRNA expressions without any prior knowledge.

First, they have prepared their data ready to use by performing various preprocessing processes on the data like discretization. Subsequently, miRNA-mRNA pairs were filtered using a threshold to get more accurate results. Then, these pairs have been classified and separated from each other. Afterwards, the Jaccard Index, which is a statistical value, was calculated and thus differentially expressed miRNA-mRNA pairs were determined. Finally, the final miRNA-mRNA modules were obtained by grouping these pairs with another statistical calculation.

2.5 Probability Learning Based Techniques

Another method used to generate miRNA - mRNA groups is the probability learning based techniques. With this technique, interaction probabilities of known miRNA-mRNA pairs are estimated. It is critical that this operation be performed robustly and effectively. Therefore, having more than one source of information is indicated as a more accurate approach. In this study [33], the idea that miRNAs and mRNAs, which have similar biological functions, should be in the same group while creating miRNA-mRNA groups. The data used are the expression data of both molecules namely miRNA and mRNA. In addition to these data, they also used predicted miRNA-mRNA pairs. However, when looking at the miRNA and mRNA expression data, it was observed that the values were generally in very different ranges. Therefore, it is thought that directly combining these two datasets will prevent the real information from being obtained. At this point, they used a population-based probability learning algorithm using different learning and probability algorithms [34–36]. In this way, they claimed that they combined different datasets in a much more robust way. The solution approach followed by the authors in this study are as follows. Initially, both miRNAs and mRNAs were randomly selected to create random populations.

Vector populations were created to express each population in 1D dimension. In the 2nd step, a score called fitness measurement is calculated for each element in both populations. In the 3rd step, the element with the best fitness score is determined, and thus the starting point for the functionalization of the co-evolutionary learning algorithm is determined. Step 4 is the update of the probability vectors. Two values specified in the study were used to update these probabilities. As stated in the study, the closeness of these values to 0 indicates that both probability vectors are very dependent on each other. In step 5, new populations are generated by updating probability distributions and this operation will continue till the condition met.

2.6 Matrix Decomposition Based Techniques

Another important technique used in studies in this field is the Non-Negative Matrix Factorization technique. It is stated that this method is successful in the integration process by successfully separating different information sources. In this study [37], the integration of information obtained from different sources was carried out successfully and miRNA-mRNA groups were formed. In some previous studies for this purpose, successful results were not obtained in the formation of miRNA-mRNA groups . Penalties are included in the equation in this method called SNMNMF to solve the problem in this study. In this way, not only this problem that caused the problem was eliminated, but also the interpretability of the miRNA-mRNA groups obtained was increased. It is not possible to obtain the optimum result by using classical methods in NMF problems. For this reason, the process is iteratively progressed in SNMNMF. In this study, an objective function with three members is created to obtain the local minimum. The first of these members deals with expression matrices, while the second and third members deal with the constraints caused by miRNA-target pairs. The solution approach of SNMNMF is as follows. This method needs some inputs to work. These inputs consist of 4 matrices. The first two of these matrices are the matrices that hold the expression information. Another matrix is the matrix where protein - protein interactions are kept. The last matrix is the matrix where the miRNA-target interactions are kept. By factoring the expression matrices, a new W matrix and two matrices in which the coefficients are held are produced. In addition, constraints are created using these two matrices. Using all of these produced matrices, miRNA-mRNA groups were formed.

2.7 Rule Induction Based Techniques

Another technique used for the detection of miRNA-mRNA groups is rule induction-based techniques. The rules produced in these techniques are based on information theory. Generally, in this technique, as in other techniques, data obtained from more than one data source is tried to be integrated.

In this study [38], a rule induction-based technique was used to find miRNA-mRNA groups. The rule generation system used to generate the rules in this study is the CN2-SD system [39]. Two different methods are used in studies aiming to generate rules for the formation of miRNA-mRNA groups. The first of these methods is the exhaustive search method, which is the easiest but has the highest computational cost. Another method is the divide and conquer method. The solution approach of this study as follows. To begin with, after using the correlation information to find similarities between genes, the genes have been divided into 2 classes namely similarity and dissimilarity. In addition to this information, class information was added using miRNA-target information data. Afterwards, rules were created using the CN2-SD rule generation system. Then, all of these rules were filtered by eliminating the rules that were found in the acquired rules and were considered to be unimportant.

Another methodology used in rule-based systems is the method of creating rules by clustering. In this methodology, it is aimed to detect similar biomarkers namely miRNAs and mRNAs. In this study [40], determining the relationships between miRNA and mRNAs was prioritized using data obtained from cancer patients. The solution approach of this study is as follows. To begin with, miRNA expression data are aggregated to form miRNA groups. It has been investigated whether it can distinguish case and control classes by using SVM algorithm in these clusters. Different validation techniques were used to clarify whether the results obtained were valid or not, and it was seen that the results produced by the obtained groups were valid. Then, using miRNA - target information, mRNA targets were added to the rules and with this way mRNA rules were generated. Using miRNA- target information is the key to reduce the number of rules obtained according to the authors. Afterwards, mRNA rules were produced using the RH-SAC algorithm. In the final step, as a result of combining miRNA and mRNA rules, miRNA - mRNA groups were generated.

2.8 Existing Tools

To the best of our knowledge there are few tools which is doing classification using miRNA-mRNA modules. These tools are namely SVM-RCE [41], SVM-RCE-R [42], and maTE [11]. SVM-RCE is a bioinformatics tool that classifies using only mRNA data. In this context, it firstly detects the gene groups with the K-Means algorithm. SVM-RCE uses correlation information for generating mRNA clusters. The performances of the generated clusters are evaluated with a ranking strategy in each iteration. As a result of this process, it is aimed to increase the classification success by removing clusters that have a low contribution to the classification problem. These clusters are merge to 1 cluster again, and the next iteration is passed with the more relevant genes. As mentioned before, SVM-RCE is an iterative approach. With this iterative approach, it is aimed to increase the classification success in each step. This process continues until the desired number of clusters is reduced. In each iteration, the classification process is tested on the data reserved for testing, and performance metrics are obtained.

One of the other existing bioinformatic tool is maTE. With this tool developed by authors, they aimed to separate patient samples from control samples in the best way by using integrated miRNA and mRNA data. Machine learning algorithms has been used for this classification problem. This tool is not only find genes that are effective for the disease but also finds miRNAs which related with these genes. These obtained miRNA-mRNA relationship information are grouped over miRNAs and with this way clusters are obtained. Acquired clusters are subjected to a ranking algorithm to find groups that can better separate patient and control samples in other words control vs case. The general logic of the algorithm is as follows. First, they performed various preprocessing operations such as normalization on the data. Subsequently, they identified the genes that were valued to be statistically significant and whose number could be a maximum of 2000. Then, data was divided into two as a train and test set and used input for Random Forest algorithm. This process continued throughout the iteration they specified, and the clusters created in each iteration were ranked. In addition, the performance results obtained by determining the best groups in each iteration are also reported

Chapter 3

Materials & Methods

In this section, everything used in this thesis will be explained within the scope of the materials and methods. In terms of materials, miRNA and mRNA datasets, statistical metrics for normalization, gene selection, clustering of interactions, classification algorithm which used to separate case and control classes, development environments will be mentioned. In terms of methods, the solution approaches and implementations of the two developed tools will be mentioned.

3.1 Datasets

The data used in this thesis has been downloaded from The Cancer Genome Atlas (TCGA) data repository [43]. There are approximately 20,000 samples of 33 cancer types in the TCGA repository [44]. In addition, this portal also contains different -omics data and the size of this data is approximately 2.5 petabytes. 11 of the 33 cancer types were used in this thesis. The TCGA abbreviations of these cancer types are: Urothelial Bladder Carcinoma (BLCA), Breast Invasive Carcinoma (BRCA), Kidney Chromophobe (KICH), Kidney Renal Clear Cell Carcinoma (KIRC), Kidney Renal Papillary Cell Carcinoma (KIRP), Lung Adenocarcinoma (LUAD), Lung Squamous Carcinoma (LUSC), Prostate Adenocarcinoma (PRAD), Stomach Adenocarcinoma (STAD), Thyroid Carcinoma (THCA) and Uterine Corpus Endometrial Carcinoma (UCEC). The number of cases and deaths in both used and unused data sets are shown in Figure 3.1. When looking at Figure 3.1, although the number of cases and deaths seem almost equal, used cancer types were selected from the more common types. This utilized datasets details are shown in Figure 3.2. Additionally, summary of the data used can be seen in Table 3.1. As mentioned before, both miRNA and mRNA sequence data were used. Z-score normalization process have been carried out on the downloaded raw data. Data was converted into RPM for miRNA-seq data and RPKM for mRNA-seq

data. In the literature, a threshold is used in RNA-seq data to differentiate cancer cells from non-cancerous cells [45]. This usage has been widely used in the literature and has been standardized. Within the scope of this thesis, to be able to select miRNAs and mRNAs for differential expression, it has been decided that this number should be at least 0.5.

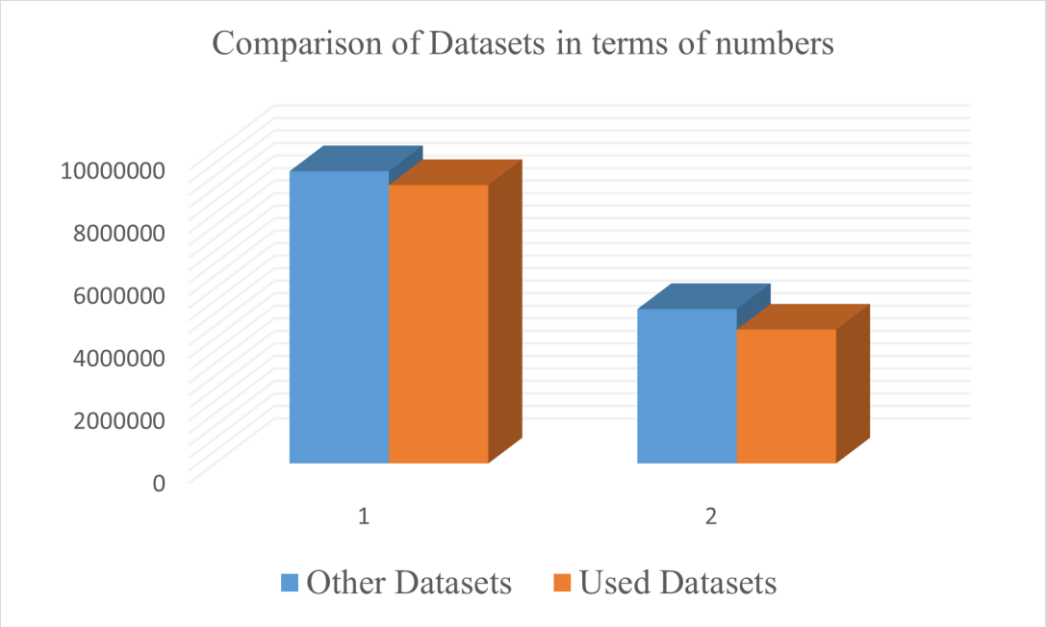

Figure 3.1 Case and Death Numbers in Both Used and Unused Datasets

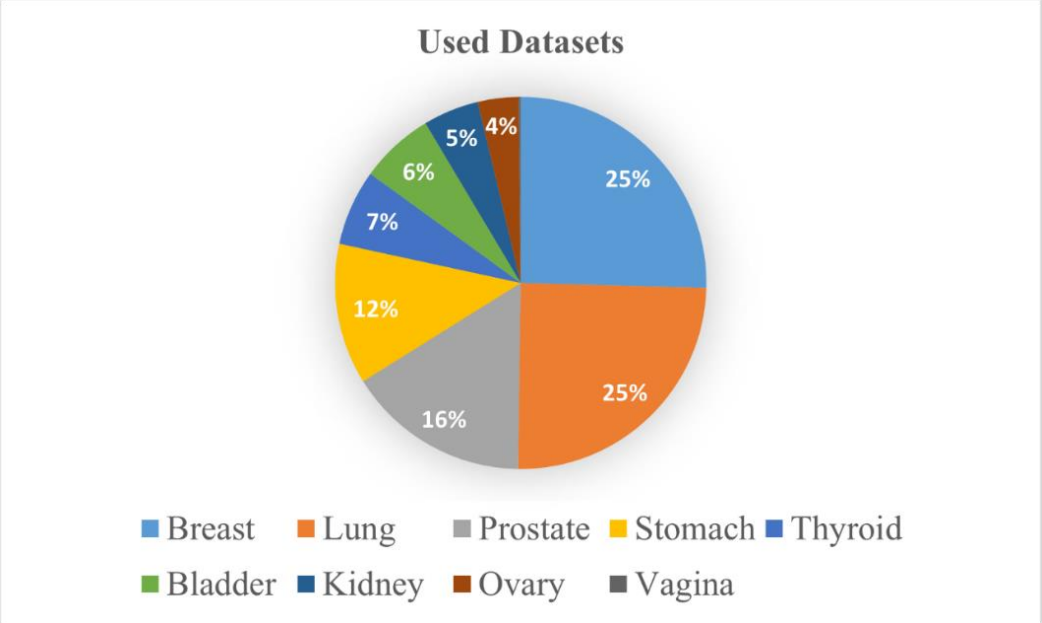

Figure 3.2 Details of Utilized Datasets

Table 3.1 Utilized Datasets Details

Disease Name	Normal Cases	Disease Cases	Pubmed ID
BLCA	405	19	24476821 [46]
BRCA	760	87	31878981 [47]
KICH	66	25	25155756 [48]
KIRP	290	32	28780132 [49]
KIRC	255	71	23792563 [50]
LUAD	449	20	25079552 [51]
LUSC	342	38	22960745 [52]
PRAD	493	52	26544944 [53]
STAD	370	35	25079317 [54]
THCA	504	59	25417114 [55]
UCEC	174	23	23636398 [56]

3.2 Statistical Metrics

3.2.1 Z-Score Normalization

Nowadays, machine learning algorithms are used in classification problems. These algorithms use the properties and values, in short, data, which are given to them as input to perform their executions. However, the ranges, types and structures of the data obtained today may differ. This causes the algorithms to work incorrectly and thus produce incorrect results. In order to prevent this situation, normalization processes are used to put the data into similar forms. Within the scope of this thesis, Z-Score normalization was used to normalize the data in different intervals. The formula that expresses how the Z-Score normalization is done is shown in the Eqs. (3.1).

$$\text{Normalized}_{\text{Value}} = \frac{\text{Observed}_{\text{Value}} - \mu}{\sigma} \quad \textbf{(3.1)}$$

In this equation, μ refers to the average value of the column that is, the relevant feature in the data. σ refers to the standard deviation. According to this formula new value will not be in a specific range. New values could be negative, 0, or positive. If the data is less than the average of the relevant column, it gets a negative value, if it is larger, it gets a positive value. In case of equality, the new value will be the value 0.

3.2.2 T-test

T-test is a statistic used to determine whether there is a statistical difference between two classes based on various assumptions. These assumptions are that each feature that constitutes the data is independent from each other, the data has a normal distribution, and finally, the variance is found homogeneously in both classes. T-test can be performed in 3 different ways, depending on the data to be used. If there is data for only one class, then paired t-test is used. If there are two different classes, then an independent t-test is used. Finally, one-sample t-test is used to compare a class with a value. T test formula is shown in Eqs. (3.2).

$$t = \frac{Grp_1 - Grp_2}{\sqrt{s(\frac{1}{n_1} + \frac{1}{n_2})}} \tag{3.2}$$

In this formulation, Grp_1 and Grp_2 indicate the average values of the respective classes compared, s_2 denotes the standard error, while n_1 and n_2 indicate the number of samples in the classes. Within the scope of this thesis, in order to obtain genes that are important for the disease being studied, t-test was used. With this way the most important genes were determined which could be 2000 as maximum.

3.2.3 Pearson Correlation Coefficient

Pearson correlation coefficient (PCC) metric is also a statistical metric used to show the presence of correlation between features, in other words, the relationship between features by using some assumptions. The assumption that properties have a linear relationship with each other is the most critical assumption. Another assumption is that the samples are independent from each other. The formulation used in PCC calculation is shown in Eqs. (3.3).

$$PCC_{mk} = \frac{\sum_{i=1}^{n}(m_i - \overline{m})(k_i - \overline{k})}{(n-1)s_m s_k} \quad \textbf{(3.3)}$$

In this formulation, while m and k express data sets, n is the total number of objects, s_m and s_k are standard deviations. $\overline{m}$ and $\overline{k}$ are the arithmetic mean of the data sets. Within the scope of this thesis, PCC is one of the two metrics used in the detection of related miRNA and mRNA pairs. Additionally, PCC it has also been used in the generation of miRNA and mRNA clusters.

3.2.4 Mutual Information

Mutual information (MI) is an entropy-based statistical metric used to measure the relationship between two features. It can be used to understand how much information a feature carries about the other one. In other words, it provides an idea whether the two features are related or not. Mutual information formulation is given in Eqs. (3.4).

$$MI(Mi, M) = \sum_{mi \in Mi} \sum_{m \in M} P(mi, m) \log \frac{P(mi, m)}{P(mi)P(m)} \quad \textbf{(3.4)}$$

In this formulation, in order to find the MI between miRNA and mRNA, Mi represents a miRNA, M represent a mRNA and P(Mi, M) represents the joint distribution of these of Mi and M. Moreover, P(Mi) and P(M) represent marginal distributions. Mutual information has been used within the scope of this thesis to find related miRNA-mRNA pairs, just like PCC. In addition, it has been used in the generation of miRNA - mRNA groups through the use of a threshold. Using two different metrics, it was also aimed to compare the performances of groups generated with two different metrics for the classification problem.

3.3 Classification Problems and Random Forest Algorithm

Classification problems have an important place in the field of machine learning. According to the criticality of the study, the importance of correctly classifying the samples increases significantly. For example, as mentioned in this study [57], the

classification problem of credit card transactions is very critical. Because transactions classified as fraudulent transactions are very limited compared to genuine transactions, but they can have very dramatic financial effects. Similarly, classification of cancer patients and cancer free patients is also very critical, because even one death poses a serious problem in healthcare practice. In addition, the direct effect of this classification success on treatment approaches should not be ignored.

In this thesis, it is aimed to solve the problem of classifying cancer patients and patients without cancer. Since this problem includes two classes, it is a binary classification problem. Random Forest algorithm is used to solve this problem. Very effective results have been obtained with the random forest algorithm. Thus, other state-of-the-art machine learning algorithms have not been used. Random forest algorithm is a supervised machine learning algorithm. In other words, it uses the previously acquired class label information. As concept, this algorithm can be used in both classification and regression problems. Random forest algorithm is a two-step algorithm. While a random forest is generated in the first stage, classification is performed by creating a model in the second stage. In the first stage, random forest algorithm generates a large amount of decision trees in a random way using the training dataset. In this way, not only the robustness is increased by producing an ensemble method but also the overfitting problem, which is the most important shortcoming of the decision tree algorithm, is tried to be avoided. In the second stage, each decision tree in the random forest makes a classification. All these classifications are evaluated using the majority voting method, and thus the classification result of the model is determined.

3.4 Feature Selection Methodologies

Some strategies are applied for the decision trees used in the algorithm to achieve the optimum classification success. These strategies are used to find answers to questions such as which feature will be on the root node of the tree, which feature will be located where in the tree, or how the tree will be divided into branches. Various metrics are used in order to find answers to these questions. These feature selection methodologies are information gain (IG), information gain ratio (IGR) and Gini index (GI), respectively. The formulations of these metrics are shown in (3.5, 3.6 and 3.7).

3.4.1 Information Gain

Information gain is a metric used to increase classification success. Information gain is an entropy-based metric. Entropy is a concept that allows the randomness of data to be understood. Its formulation is shown in Eqs. (3.5).

$$\text{Entropy}(X) = \sum_{i=1}^{c} -p_i \log_2 p_i \quad \textbf{(3.5)}$$

In this formulation p_i denotes the probability of class i. In this formulation, the entropy value should be minimized in order to achieve maximum classification success. In this way, the most informative features are obtained, and trees are generated according to this information. The formulation of IG for a feature (Y) is shown in Eqs. (3.6).

$$\text{IG}(Y) = \text{Entropy}(\text{Dataset}) - \text{Entropy}(Y) \quad \textbf{(3.6)}$$

3.4.2 Information Gain Ratio

Information Gain Ratio is a more specialized version of the IG metric. The most important shortcoming of the information gain metric is that it tends to select the feature that has more different data in it. This situation may cause overfitting. This metric has been introduced in order to reduce the effect of these drawbacks to some extent. IGR tries to normalize the erroneous results that can be obtained with IG by using underlying information and thus overfitting is tried to be avoided. In order to calculate IGR, underlying info must be calculated first. The formulation used to compute this metric for a feature (Y_i) against its class (M_i) is shown in Eqs. (3.7). The formulation of IGR is shown in Eqs. (3.8).

$$\text{Underlying\_Information}(Y, Mi) = -\sum \frac{|Y_i|}{|Y|} \log_2 \frac{|Y_i|}{|Y|} \quad \textbf{(3.7)}$$

$$\text{IGR}(Y, Mi) = \frac{\text{IG}(Y, Mi)}{\text{Underlying\_Information}(Y, Mi)} \quad \textbf{(3.8)}$$

3.4.3 Gini Index

Gini index formulation is shown in Eqs. (3.9). As can be seen from Gini index formulation, it is a metric that expresses the probability of a randomly selected feature being classified incorrectly. The value that this index can take is the range of 0 to 1. As can be understood from this statement, features with a minimum GI value are preferred when generating the decision tree.

$$GI = 1 - \sum_{i=1}^{n} (P_i)^2 \quad \textbf{(3.9)}$$

In this formulation n refers to number of classes and P_i refers to probability of a feature classified for a separate class.

3.5 Platforms

In this section, the platforms used while implementing the two tools developed within the scope of this thesis will be discussed.

3.5.1 KNIME Analytics Platform

KNIME analytics platform is an open-source data science tool that is constantly developed and updated [58]. By using KNIME, workflows that perform various tasks can be created. In addition, data extraction operations are carried out by connecting with many databases. Due to the ability to connect to web servers, it also enables to get and post requests and to use the responses. A wide range of preprocessing operations like data cleaning, data normalization, data discretization can be easily performed on the data and the obtained findings can be easily visualized. Also, machine learning and deep learning applications can be performed easily. In addition to all these, thanks to all third-party plugins developed by developers, end users can execute very specific tasks. Usage of these third-party plugins increases the power of KNIME in terms of functionally This approach allows complex and time-consuming tasks to be completed in a very short time. The interface of the KNIME analytics platform is shown in Figure 3.3.

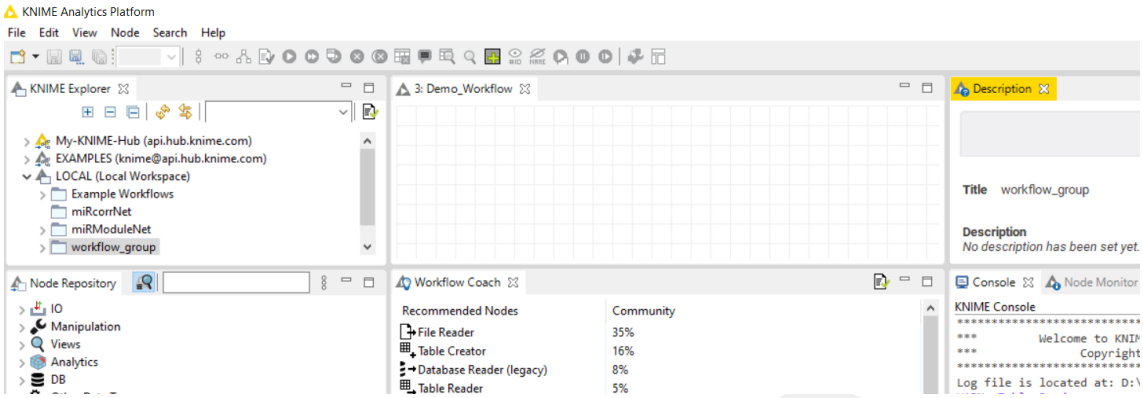

Figure 3.3 Graphical User Interface of KNIME

As can be seen from Figure 3.3, KNIME is very easy to use. In Figure 3.3, there is an editor where the workflows are developed, the Solution Explorer section with all the developed solutions can be seen in it, the node repository section where we can search the nodes, and the Workflow Coach section that makes node-based suggestions in the workflow. There is also a description section where people who will use workflow can get an idea about workflow. The operations that are desired to be carried out in the KNIME analytics platform, performed by using workflows. In the workflow, there are structures called nodes, each of which has a special task. Tasks are performed using these nodes. KNIME also supports node types such as metanode and component in order to increase Encapsulation, that is to prevent the end user from getting lost in detail.

3.5.2 Other Platforms

In addition to the KNIME platform, another environment used during development is the R environment [59]. R is an open-source programming language. In addition to performing statistical operations, R is also used in visualization operations. Due to the publicly availability of the R platform, a serious community has been formed. In this way, libraries that can perform a wide variety of tasks have been produced and used today. R can work integrated in a structure that can receive and respond to queries from the KNIME environment. Within the scope of this thesis, not only differentially expressed mRNAs and miRNAs were detected but also Mutual Information values were calculated with R. The last different environment used in these implementations is the Python programming language. Python can run as embedded in

KNIME. In this way, scripts developed with Python programming language were used among the tools developed within the scope of this thesis.

3.6 Proposed Method : miRcorrNet

In this section, miRcorrNet, the first of the two tools developed within the scope of thesis studies, will be explained. miRcorrNet stands for miRNA-mRNA Correlation Network. miRcorrNet has been developed based on previously developed SVM-RCE and maTE tools. This general flowchart and its components are shown in Figure 3.4.

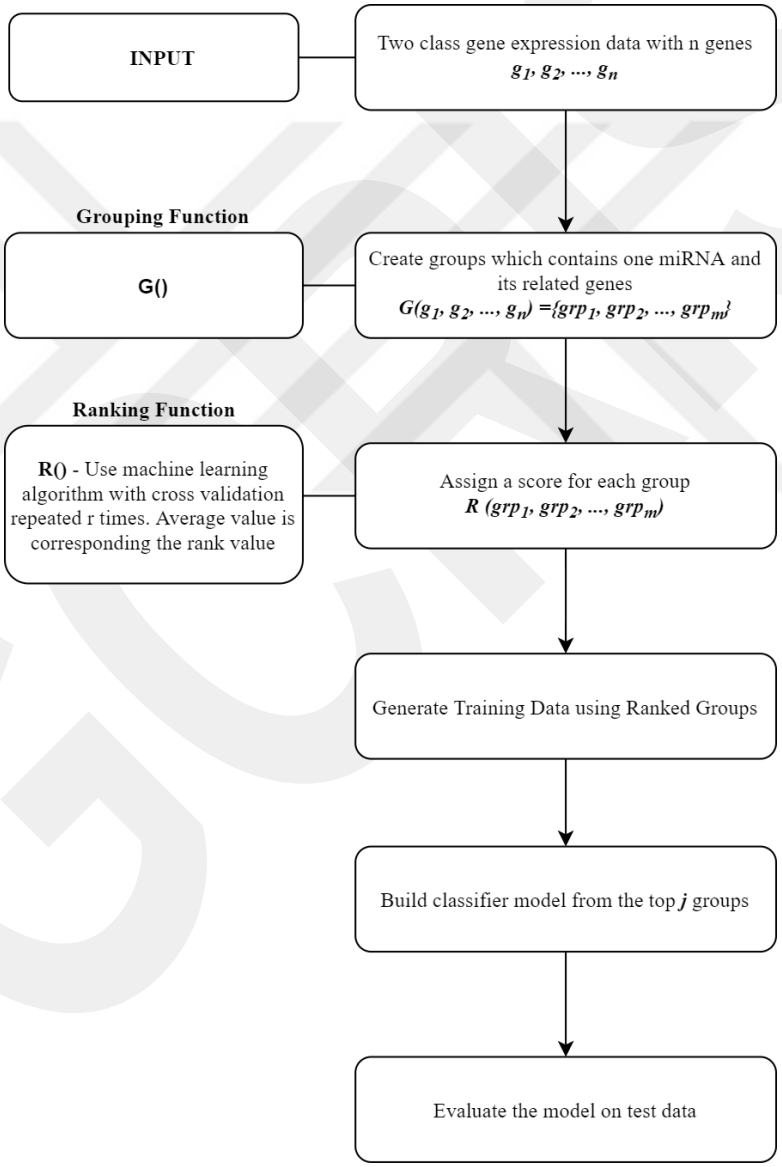

Figure 3.4 The General Used Approach

The critical number of elements in this general approach is two. One of them is G() , the Grouping Function, which is used for grouping in other words used for generating clusters which contains a single miRNA and related mRNAs. The other element is R(), the Ranking Function, which aims to classify the obtained groups using a score to be generated.

3.6.1 Grouping Element

Here, the G() function may vary according to the approach of the method desired to be used. There are 3 different options for this function. As a first option, this function can be any of the state-of-the-art computational clustering methods such as K-Means used in SVM-RCE-R. As a second option, this function can also be generated using biological information-based databases that are used for biological target validation in the literature. As a last option, this function can also be generated in a hybrid way, combining these two methods mentioned earlier. Some chunk of the output of this G() function generated using Thyroid Carcinoma data is given in Table 3.2.

Table 3.2 Some Chunk of the G() Function Output Using THCA Data

miRNA	**Gene lists associated with a specific miRNA**
hsa-miR-222-3p	*LRP1B, PLA2R1*
hsa-miR-652-3p	*TMEM41A*
hsa-miR-497-5p	*TMEM41A*
hsa-miR-139-5p	*ERBB3, RUNX1*
hsa-miR-152	*C11orf80, ARHGAP23, GALNT7, PLCD3, QPCT, LAMB3, ICAM1, ABTB2, SPATS2L, MCTP2, MET, SYTL1, BID, PIAS3*
hsa-miR-30c-2-3p	*MTMR11, RUNX1*
hsa-miR-28-5p	*GALNT7, EPS8, CTTN, TMEM41A, IGF2BP2, KIAA0284*
hsa-miR-28-3p	*GALNT7, PSD3, KLHL2, MCTP2, GPRC5B, TUSC3 ,POU2F3, GALE, SLC35F2, P4HA2, EPS8, MEAF6, SDC4, CTTN*

hsa-miR-126-5p	*PLCD3, ERBB3, MACC1, TBC1D2, SLC34A2, MET, CDC42EP3, SPTBN2, LAD1, AMOT, PROS1, PVRL4, SLC35F2, PERP*
hsa-miR-148b-3p	*ARHGAP23, PLCD3, DOCK9, ERBB3, FAM20C, ICAM1, ABTB2, MICAL2, TBC1D2, SLC34A2, MET, BID, KCNQ3, LGALS3, SPTBN2, LAD1, MTMR11, FAM176A, RARG, ETHE1*
hsa-miR-195-5p	*GALNT7, SPINT1, TMEM41A, IGF2BP2*
hsa-miR-30a-5p	*C11orf80, LAMB3, BID, MTMR11 ,RUNX1, FLJ42709, CDC42EP1, NFE2L3, CD276, RUNX2*

The groups shown in Table 3.2 can be generated with the 3 methods mentioned before. In this thesis, these groups were generated with statistical metrics like PCC and MI based on computing, which is the first option. Each group in this table consists of a miRNA and genes which thought to be potentially related to this miRNA. The potential existence of this relationship is dependent on a threshold. When we look at the table, let us consider *hsa-miR-28-5p* as an example. The group with this miRNA includes *GALNT7*, *EPS8*, *CTTN*, *TMEM41A*, *IGF2BP2* and *KIAA0284* genes as related mRNAs. Thus, these genes can be potential target genes of the respective miRNA. Once groups are generated, the most important issue is to determine the effects of these groups on the classification performance. Ranking Function i.e., R(), which is the second critical element in the general approach, is used to make this evaluation.

3.6.2 Ranking Element

The pseudocode of the algorithm used to generate a score for each group is given in Figure 3.5. As a result of this process, each group will have a score that shows the ability to separate two classes. Obviously, machine learning algorithm is used for this ranking process. At the end of this R() function's execution there will be an output which contains a list of groups in an ascending order in terms of score. Afterwards, the groups that are planned to be used in the classification phase are determined. This selection can be made using individual groups or a certain number of predefined groups, or it can be achieved by using a certain threshold.

Ranking Algorithm -R(X_s, M, *f*, *r*)

$\mathbf{X_s}$: any subset of the input gene expression data X, the features are gene expression values

M {is a list of groups produced G() function}

f is a scalar: split into train and test data

r: repeated times (iteration)

res={} for aggregation the scores for each m_i

Generate Score for each m_i *and then rank according to the score, Rank(m_i):*

For each m_i in M

sm_i=0;

Perform *r* time (here r=5) steps 1-5:

1. Perform stratified random sampling to split X_s into train X_t and test X_v data sets according to *f* (here 80:20)
2. Remove all genes (features) from X_t and X_v which are not in the group m_i (Creat sub data that contains just the genes belongs to group m_i)
3. Train classifier on X_t using SVM
4. t = Test classifier on X_v –calculate performance
5. $sm_i = sm_i + t$;

$Score(m_i) = sm_i / r$; Aggregate performance

res= { Union of Score(mi) }

Output

Return {$Rank(m_1)$, $Rank(m_2)$, …, $Rank(m_p)$} which is the sort of the list based on the score value of each group

Figure 3.5 Ranking Algorithm for Acquired miRNA-mRNA Groups

3.6.3 miRcorrNet Tool

Two different -omics data are used in miRcorrNet tool, namely mRNA expression and miRNA expression. As notation, T_{mRNA} denotes the mRNA expression and T_{miRNA} denotes the miRNA expression data. Both data are in 2D structure, that is, data consisting of rows and columns and belonging to the same individual. Table 3.3 presents the representation of this data in the form of a table.

Table 3.3 Sample Expression Data for miRcorrNet

mRNA Expression Data

Sample ID	Class	A1BG	A2LD1	…	ZZZ3
TCGA-DK-A6AV	neg	32.877	28.283	…	721.166
TCGA-DK-A3WX	neg	39.634	57.526	…	593.293
TCGA-GC-A3WC	pos	29.789	98.344	…	1,057.069
TCGA-BT-A20N	pos	37.378	55.011	…	755.688
…	…	…	…	…	…

miRNA Expression Data

Sample ID	Class	hsa-let-7a-3p	hsa-miR-7a-5p	…	hsa-miR-99b-5p
TCGA-DK-A6AV	neg	44.775623	13345.98449	…	9772.686386
TCGA-DK-A3WX	neg	34.30313	17531.35061	…	13508.08329
TCGA-GC-A3WC	pos	9.389	15229.41331	…	18601.78121
TCGA-BT-A20N	pos	3.534104	4717.325745	…	6845.372094
…	…	…	…	…	…

The workflow of miRcorrNet tool is shown as a flowchart in Figure 3.6. There are basically 9 steps in this workflow. Two of these 9 steps are the grouping and ranking functions as mentioned earlier. While taking the input as raw data is the starting point of the workflow, the flow is completed by evaluating the performance results obtained from the model produced by the machine learning algorithm.

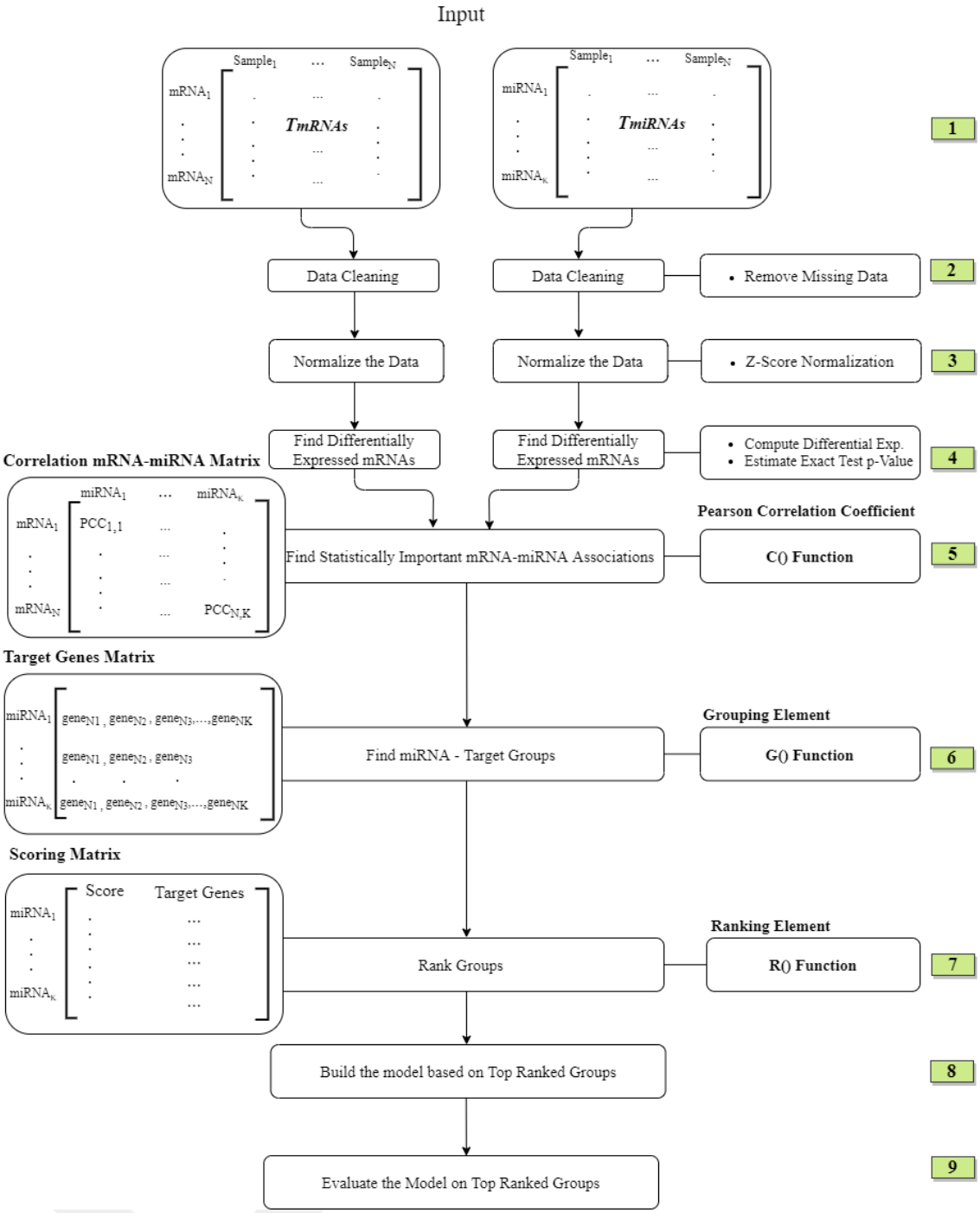

Figure 3.6 Solution Approach of miRcorrNet

In 1st step, mRNA and mRNA expression data are uploaded after preparing according to Table 3.3. The missing data problem, which is frequently encountered in today's real-world data, is dealt with in the 2nd step. Data cleaning is performed by removing columns containing missing data. Following the cleaning of the missing data, the data normalization phase, which is a part of the preprocessing step, has been started.

Due to the continuous nature of the expression data, all data were normalized using Z-Score normalization method. With the completion of the preprocessing step, the data is ready for use. In the 4^{th} step, mRNAs and miRNAs that could potentially be important in the context of the current studied disease were identified using the t-test method. During this process, the threshold, which indicates the statistical significance, was set as 0.05. Then, Pearson Correlation Coefficient (PCC) was used to detect the associations between these mRNAs and miRNAs. As mentioned earlier, miRNAs tend to have a negative correlation on mRNAs. In this context, the pairs below a certain threshold were used. This threshold is chosen arbitrarily as -0.6, but this value can be changed by end users. Then, using all this information, a 2D mRNA-miRNA correlation matrix was created. This process is shown as the C() Function. The next step is to create mRNA-mRNA groups previously expressed as a G() function i.e., Grouping element. Groups were generated from the idea that mRNAs associated with the same miRNA should be a group. In this way, the target genes matrix was created. As can be seen from Figure 3.6, the number of mRNAs that each miRNA can be associated with can vary. In the 7^{th} step, these groups were subjected to a ranking process in order to determine their contribution to the solution of the classification problem. This step corresponds to the R() function i.e., Ranking Element. Details of the Ranking Element are shown in Figure 3.7. In the 8^{th} step, a machine learning model was created using groups and / or groups whose contributions were considered to be high in terms of pValue. For miRcorrNet a predefined number of groups were used, and this number was 10. In the last step, by looking at the results obtained by running the KNIME workflow, the 10 groups decided to be used were evaluated and compared according to various performance metrics.

The Ranking element shown in Figure 3.7. This element is used to obtain the performances of the produced groups. One example of this element's output is given in Table 3.4. This element uses T_{mRNA} and the generated target genes matrix as input. Sub-datasets are generated by adding class information to the group of miRNA and related mRNAs which is located in each line in the target genes matrix. Afterwards, these sub-datasets produced are forwarded to the next element to calculate their score by the cross-validation method. Random Forest algorithm was used for this process. Finally, the scores of each miRNA-mRNA group were obtained by keeping the score in the scoring matrix as its output.

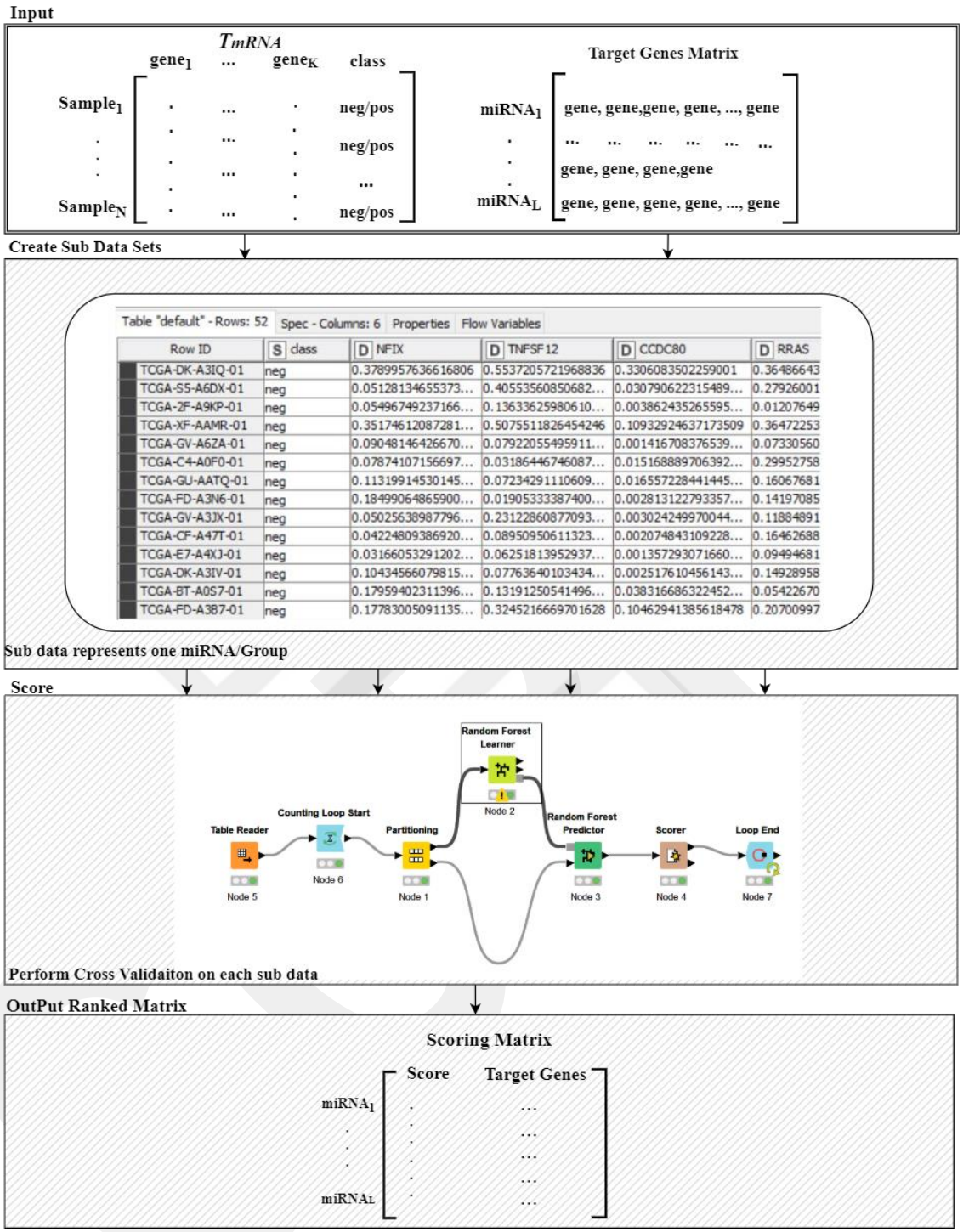

Figure 3.7 Schematic Representation of the Ranking Function

Table 3.4 Example Output of Ranking Operation Applied on BLCA Data

Cluster/Group	ACC	SEN	SPE	Recall	PRE	F-Measure	Cohen's Kappa
hsa-miR-32-5p	0.65	0.55	0.71	0.55	0.61	0.52	0.27
hsa-miR-361-3p	0.85	0.70	0.94	0.70	0.87	0.76	0.66
hsa-miR-205-5p	0.91	0.90	0.91	0.90	0.86	0.88	0.81
hsa-miR-30e-5p	0.76	0.60	0.86	0.60	0.77	0.60	0.46
hsa-miR-181a-5p	0.89	0.75	0.97	0.75	0.96	0.82	0.75
hsa-miR-106b-5p	0.93	0.85	0.97	0.85	0.96	0.88	0.83
hsa-let-7a-5p	0.78	0.65	0.86	0.65	0.75	0.68	0.52
hsa-miR-22-3p	0.95	1.00	0.91	1.00	0.89	0.94	0.89
hsa-miR-17-3p	0.91	0.80	0.97	0.80	0.96	0.85	0.79
hsa-miR-151a-5p	0.82	0.70	0.89	0.70	0.86	0.73	0.60
hsa-miR-374a-3p	0.69	0.55	0.77	0.55	0.57	0.55	0.32
hsa-miR-186-5p	0.84	0.75	0.89	0.75	0.83	0.78	0.65

ACC : Accuracy, SEN: Sensitivity, SPE:Specificity, PRE: Precision. Whole results for this R() output have been given as mean. The columns are the performance measurement achieved by cross-validation. The rows are the name of each group that is the miRNA. The sorted table according to Accuracy is used as the rank for each miRNA.

miRcorrNet is an iteration-based tool. Whole algorithm's operations will repeat the number of times which will be specified by the end user. In each iteration, 90% of the data is used for training and 10% for testing. In addition, the imbalance class problem can be solved with under sampling.

All these groups obtained as a result of this should be prioritized among themselves. For this purpose, RobustRankAggreg, an R package, was used in miRcorrNet tool [60]. The task of this package is to give a pValue to each group produced. In this way, prioritization was made among groups.

3.7 Proposed Method: miRMUTINet

Another bioinformatic tool developed within the scope of this thesis is miRMUTINet. The solution approach workflow of miRMUTINet tool is shown in Figure 3.8. The miRMUTINet tool consists of 6 steps. In 1st step, just like miRcorrNet, in terms of data cleaning the columns were removed which have missing values. However, continuous expression data were normalized using the Z-Score normalization technique. In 2nd step, statistically significant miRNAs and mRNAs are detected using the t-test method. In 3rd step, miRNA-mRNA pairs were determined using mutual information (MI) metric and kept in Mutual Information matrix. Infotheo, an R package, was used to calculate MI scores [61]. In this matrix, only pairs exceeding the MI threshold set by the end user are evaluated. The results presented in this thesis were produced by using a value of 0.25.

Normalized data must be discretized in order to calculate MI scores. This discretization process was carried out by column wise. In order to calculate MI scores KNIME workflow uses discretized miRNA and mRNA data as input for the detection of miRNA-mRNA pairs. The data has been discretized into 3 categories. These categories names are under-expressed, neutral, and over-expressed. The numerical values of these categories are -1, 0 and +1, respectively. Formulations used to categorize data in this way is given Eqs. (3.10). In this formulation, μ is the mean of the relevant column and σ is the standard deviation value of the relevant column. Value is the data that will be discretized.

$$\begin{aligned} &\text{Value} < \mu - \sigma \Rightarrow -1 \\ &\mu - \sigma < \text{Value} < \mu + \sigma \Rightarrow 0 \\ &\text{Value} > \mu + \sigma \Rightarrow +1 \end{aligned} \quad \textbf{(3.10)}$$

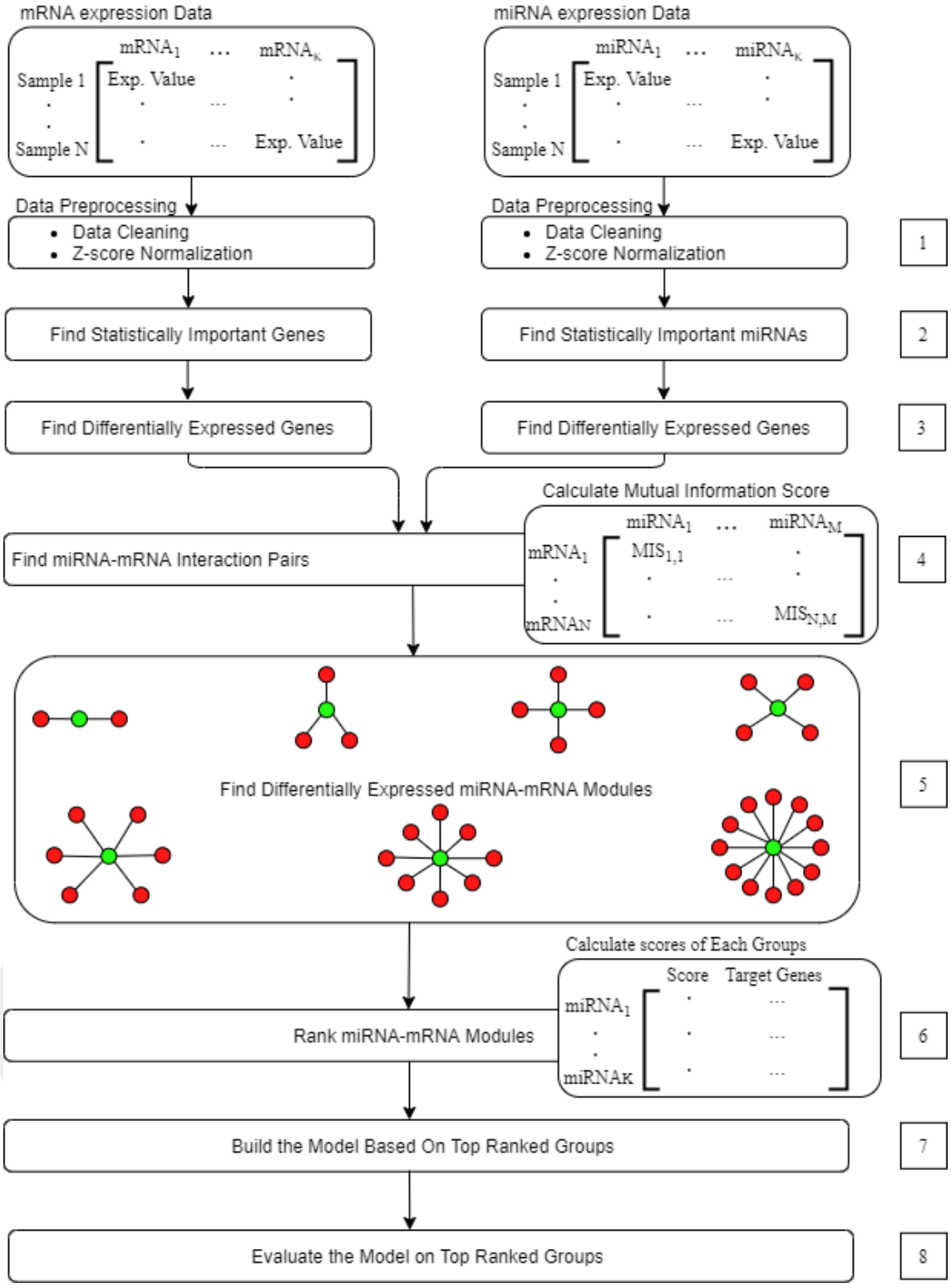

Figure 3.8 Solution Approach of miRMUTINet

In 4th step, miRNA-mRNA modules were constructed using this MI matrix. These modules contain only a single miRNA and associated mRNAs with this miRNA. An example of the produced star shaped modules is presented in Figure 3.9. First of all, the miRNA-mRNA pair with the highest MI score was determined. In this way, the miRNA of the module is determined. The mRNA interacting with this miRNA is a starting point for mRNAs to be added. Afterwards, using the MI scores between mRNAs, those whose value exceeds the threshold set by the end user are added to the relevant module. At the end of this process, the miRNA-mRNA module is obtained. These modules, ie groups, obtained in the 5th step are subjected to the Ranking process, just like as in miRcorrNet. At the end of this step, a score indicating the ability of each miRNA-mRNA group to separate classes will have been acquired. The machine learning model was established with the groups that were decided to be used in the 6th step, and finally the classification performance was evaluated.

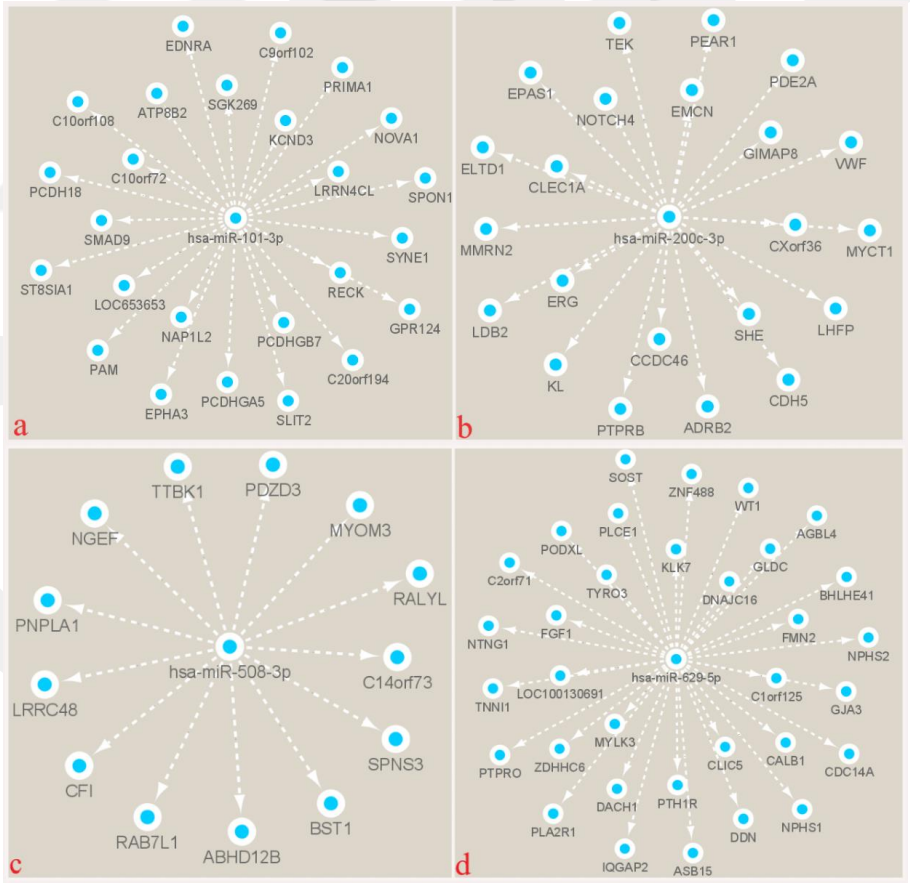

Figure 3.9 4 different examples of star shaped miRNA-mRNA modules. While the center node represents the selected miRNA, other nodes represent associated genes (mRNAs) that are found to be associated with the center miRNA

3.8 Implementation

As mentioned earlier, the main development environment in this thesis is the KNIME Analytics Platform. In this section, the KNIME workflow of the two developed tools will be shared and their details will be discussed.

3.8.1 Implementation of miRcorrNet

The final version of miRcorrNet bioinformatics tool developed in KNIME environment is shown in Figure 3.10. Although this workflow works completely automated, its operation can be expressed as follows. Using the List Files Node, referred to as node 223, the end user must specify the directory containing the previously prepared miRNA and mRNA data. The extensions of these data should also be ".table". Afterwards, Table Reader nodes obtain both mRNA and miRNA data in KNIME table form for later use. The type of the classification problem in this thesis is the binary classification problem. Thus, if there is any sample in the data except control and case classes, these examples should be eliminated by using *filtering* nodes. After these processes, both miRNA data and mRNA data are subjected to Z-Score normalization process using *Normalizer* nodes. In addition, there are 3 important parameters that end users must set in this workflow. These parameters are positive correlation threshold, negative correlation threshold, and number of iterations, respectively. The values of these parameters are set as +1, -0.6 and 100 by default. End users can change these parameters using the *SetParameters* node. These threshold values are the parameters to be used for the PCC. The Number of iterations parameter is the parameter that specifies the number of times the sequence of operations that miRcorrNet will perform. The miRcorrNet metanode, specified as node 460, is the node where the algorithm that performs all steps between 4^{th} and 9^{th} steps of miRcorrNet's solution approach runs.

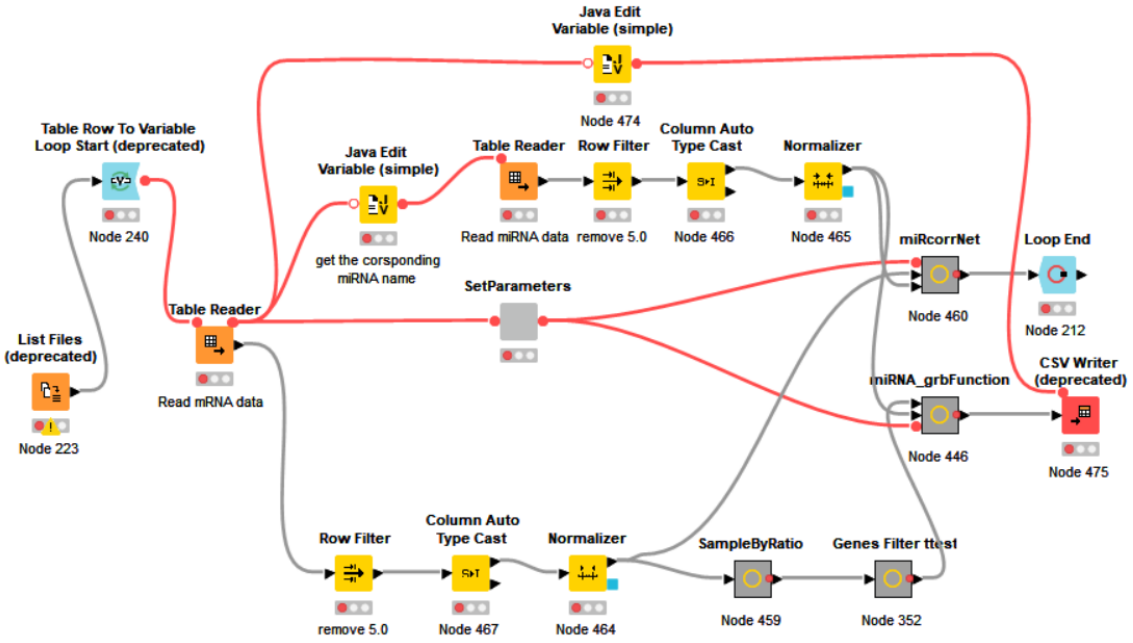

Figure 3.10 Implementation of miRcorrNet using KNIME

3.8.2 Implementation of miRMUTINet

The final version of miRMUTINet bioinformatics tool developed in KNIME environment is shown in Figure 3.11. The operation of miRMUTINet, which is similar to miRcorrNet in implementation, is as follows. In order for miRMUTINet to start working, the end user must provide the ready-to-use directory containing miRNA and mRNA data. Subsequently, these data are taken into KNIME environment and made ready for use. miRMUTINet has two parameters. These parameters are mutual information threshold and number of iterations, respectively. The numerical default value of these parameters is 0.25 and 100, respectively. These parameters can also be changed by the end user by using the *SetParameters* component node in the KNIME workflow. With the use of the *Data_Preparation* metanode, the data is tailored to the format that can be used in the miRMUTINet metanode which hosts the main algorithm. In order to generate the miRNA - target file, first, statistically significant genes are determined by the t-test method. Then the output of this node is given as an input to the *Generate_miRNA_mRNA_Modules* component node. Subsequently, the miRNA-target document is written to the specified directory through the CSV Writer node.

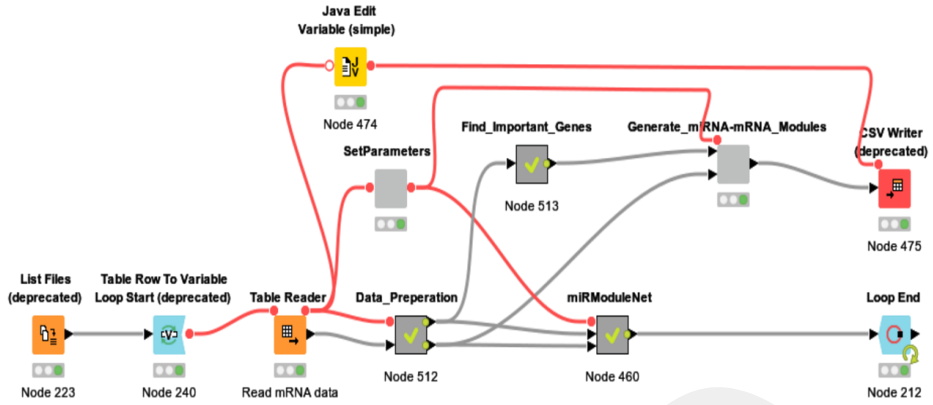

Figure 3.11 Implementation of miRMUTINet using KNIME

3.8.3 Implementation of Test Workflow

It is very important to see the results of a study as successful. In addition, the outputs obtained as a result of repeating experiments using a dataset that has never been used before are also extremely critical. In this way, it can be demonstrated whether the developed tool can produce robust and reliable results. Within the scope of the thesis, a separate KNIME workflow has been developed to carry out this process. This test workflow is presented in Figure 3.12.

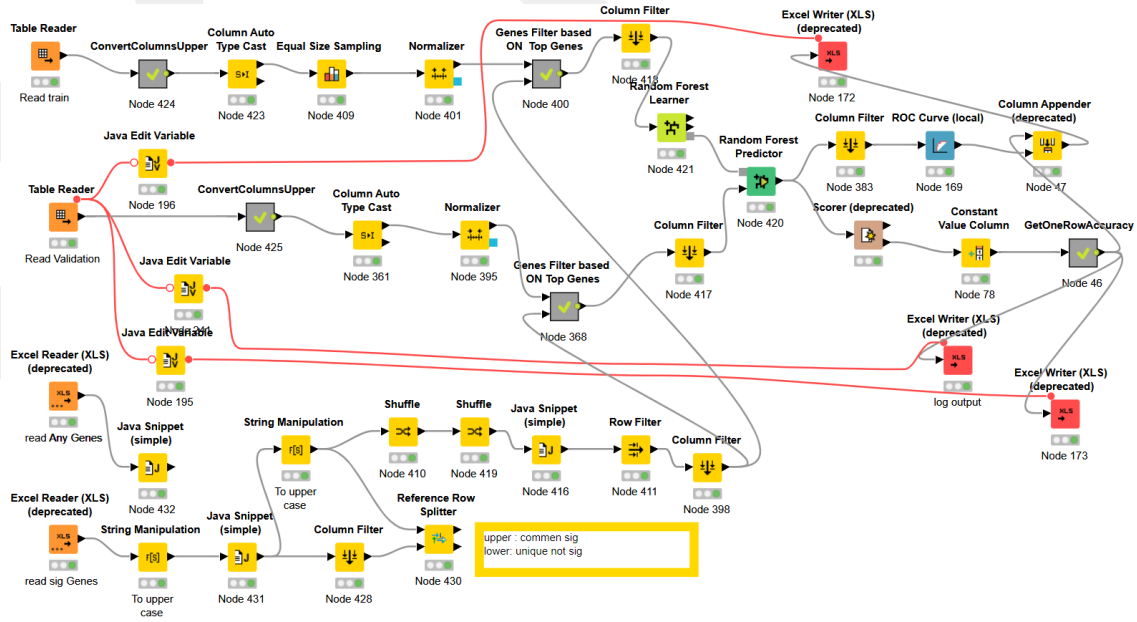

Figure 3.12 Test Workflow

As can be seen from the Figure above, this workflow requires 4 important files as input. The first of these files can be any of the data used in the development of tools for performing the train operation. Another file is the file containing the new data that has not been used at any stage of the process before. Of course, it is essential that these two data belong to the same disease in order to make logical inferences. Another file is the miRNA or mRNA data contained in the data used for the train. Random genes or miRNAs used for comparison are randomly selected from this data. The last file is the file containing the genes or miRNAs that are considered important and obtained as a result of the execution of the tool.

Chapter 4

Results

In this section, to begin with, all used performance metrics are briefly explained. Afterwards, the results obtained from the studies conducted within the scope of this thesis are given in detail. In addition, the results of all available tools have been compared using various performance metrics. Also, a pathway enrichment analysis was performed in order to understand the relationship between potentially important miRNAs found in the results and pathways, and these results were also presented.

4.1 Performance Metrics

At the most superficial terms, classification problems are divided into 2 categories as balanced and imbalanced. The success metrics used for balanced classification problems are generated by using some formulation as a result of the generation of a confusion matrix. These metrics are Accuracy, Sensitivity, Specificity, Precision, and F-measure, respectively. The confusion matrix in which these metrics are generated is given in Table 4.1. In this table, the meaning of the TP expression is that a person who is a case is determined as a case by the model. The meaning of the TN expression is the condition that someone who does not actually have a disease is determined as cancer free by the classifier model. The meaning of the expression FP is when the classifying model classifies someone who does not actually have a disease as a case. Finally, the meaning of the FN statement is the condition that a person who is case is classified as control by the classifier model. The expression that causes the greatest concern in classification problems, especially in the context of health practices, is FN. These metrics alone are not considered sufficient for imbalanced classification problems [62] . Therefore, it is thought that using the Area Under the Curve (AUC) metric may be more accurate for such problems.

Table 4.1 Confusion Matrix

		Predicted Value	
		Positive	Negative
Real Value	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Accuracy: Accuracy is a metric used to examine the accuracy of the results produced by the classifier model. Its formulation is shown in Eqs (4.1). As can be understood from the formulation, the higher the accuracy of the results produced by the model, the higher the accuracy score will be. According to the available data set, it can be evaluated whether its use will be sufficient by itself. If the available data set is imbalanced, that is, if there is a huge difference between the sample numbers of the two classes, its use alone is not sufficient. If the opposite is the case which means the class distributions are balanced, it is the most important performance metric used to evaluate classifier models.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad \textbf{(4.1)}$$

Sensitivity: This metric is a performance metric used to measure how correctly instances of the case class are predicted. The formulation of this metric is Eqs. (4.2). This metric is also called recall or True Positive Rate. It clearly measures how accurately the case samples are classified as case by the classifier model. As in this study, this metric is very critical in healthcare applications.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad \textbf{(4.2)}$$

Specificity: This metric is a performance metric used to measure how correctly instances of the control class are predicted. The formulation of this metric is Eqs. (4.3). This metric is also called False Positive Rate. It clearly measures how accurately the cancer free samples are classified as cancer free by the classifier model.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4.3}$$

Precision: Precision is a measure that shows the rate of how many of the samples estimated by the model as cases are correctly classified as cases. The formulation of precision is shown in Eqs. (4.4). This metric is an important performance metric as it shows how many percent of instances that are cases are correctly obtained by the model.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4.4}$$

F-Measure: Although the performance metrics used are calculated uniquely, this is not the case for all. There may be very dramatic differences among performance metrics calculated individually due to the used data set. One of the metrics used to detect the problems that may arise from this situation is the F-Measure metric. F-Measure is a metric obtained by calculating the harmonic mean of precision and recall metrics. It is also called F-Score or F1-Score. Formulation of F-Measure metric is shown in Eqs. (4.5).

$$\text{F} - \text{Measure} = \frac{2}{\text{precision}^{-1} + \text{recall}^{-1}} \tag{4.5}$$

4.2 Results of miRcorrNet

In the implemented workflow, the results were obtained by using the data of 11 type of cancer downloaded from the TCGA data portal. In order to obtain the results, predefined number of clusters were given, and all results were obtained between clusters 1 and 10. The number of predefined clusters used within the scope of this thesis is 10. In addition, while obtaining results in workflow, the three parameters number of iterations, positive correlation threshold and negative correlation threshold are set to

100, 1 and -0.6, respectively. An example result obtained with miRcorrNet tool is presented in Table 4.2. All the results obtained with the miRcorrNet tool are presented in Table 4.3 for clusters 1, 2, 5,7 and 10.

Table 4.2 Example of Performance Output of the miRcorrNet and miRMUTINet

#Top Groups	Number of Genes	Accuracy	Sensitivity	Specificity
10	388.79	0.94	0.92	0.95
9	355.81	0.95	0.92	0.96
8	328.58	0.94	0.91	0.96
7	288.23	0.93	0.91	0.95
6	259.99	0.94	0.92	0.95
5	223.87	0.94	0.92	0.95
4	182.58	0.94	0.91	0.95
3	146.43	0.94	0.91	0.95
2	93.16	0.93	0.9	0.94
1	45.06	0.91	0.86	0.93

This is an example of the output of the miRcorrNet, miRMUTINet, maTE, or SVM-RCE. These results acquired with miRcorrNet using BLCA data. The column #Genes is the average number of genes. In the first step. we build a model from the genes belonging to the first top group and then test it using the testing part of the data. Then we build a model from the top 1 and 2 groups then test. For j = 10. the model is built from the genes belonging to the top 10 groups and tested accordingly.

Table 4.3 Whole miRcorrNet Results

miRcorrNet Performance											
#Grp	**BLCA**	**BRCA**	**KICH**	**KIRC**	**KIRP**	**LUAD**	**LUSC**	**PRAD**	**STAD**	**THCA**	**UCEC**
10	0.98	1.00	1.00	0.99	1.00	1.00	1.00	0.95	0.96	1.00	0.99
7	0.98	1.00	1.00	0.99	1.00	1.00	1.00	0.95	0.98	1.00	0.99
5	0.99	1.00	1.00	0.99	1.00	1.00	1.00	0.96	0.97	1.00	0.99
2	0.97	1.00	1.00	0.99	1.00	1.00	1.00	0.96	0.98	1.00	0.99
1	0.97	0.99	1.00	0.99	1.00	1.00	1.00	0.95	0.93	0.99	0.99
miRcorrNet number of genes											
#Grp	**BLCA**	**BRCA**	**KICH**	**KIRC**	**KIRP**	**LUAD**	**LUSC**	**PRAD**	**STAD**	**THCA**	**UCEC**
10	407	56	4,916	245	365	352	398	122	86	278	389
7	290	60	2,998	207	316	257	270	69	52	219	269
5	211	49	2,031	162	297	181	194	54	26	173	193
2	84	32	870	70	157	65	68	21	13	92	75
1	46	24	306	35	69	29	28	10	8	48	33

Whole miRcorrNet results has shown using Area Under the Curve (AUC) value in terms of performance metric. #Grp is the number of top groups. Number of genes' mean values have been given.

4.3 Results of miRMUTINet

With the miRMUTINet tool developed within the scope of this thesis, the results were obtained by using the data of 11 different cancer data types, just like the miRcorrNet tool. The results were obtained within the predetermined number of clusters. In this context, results have been obtained for all clusters between 1 and 10. There are 2 parameters in miRMUTINet KNIME workflow that are necessary for operation. The first of these is number of parameters and its value is set to 100. The

other parameter is the mutual information score parameter. For this parameter, 3 different values, 0.15, 0.25 and 0.5, were used to compare its effect. Using these different threshold values all results have been obtained. The acquired results using these three different thresholds were compared as in Figure 4.1.

Figure 4.1 Comparison of Results Using Different Thresholds

This Figure has been prepared for all data sets, but only the figure prepared for THCA is presented here for visual purposes. In Figure 4.1, the results obtained for 3 different thresholds, namely 0.15, 0.25 and 0.5, are presented, respectively. The results presented in the Figure are presented in the order of these threshold values from top to bottom. However, considering the results obtained using the 0.25 threshold and the results obtained within the scope of indicators such as the number of genes, it was considered that it would be fairer for comparison, so it was decided to use 0.25 as the threshold value. In addition to all these, to see the effect of the number of clusters, all the results obtained with 1 cluster and 2 clusters are presented in Table 4.4.

Table 4.4 Performance Results of miRMUTINet

1 group / **Disease**	**Acc**	**Sen**	**Specificity**	**FM**	**AUC**	**Precision**	**CK**
BLCA	0.89	0.82	0.92	0.85	0.95	0.88	0.73
BRCA	0.95	0.92	0.97	0.92	0.98	0.94	0.89
KICH	0.98	0.93	1.00	0.96	0.99	1.00	0.94
KIRC	0.99	0.97	1.00	0.98	0.99	0.99	0.97
KIRP	1.00	0.99	1.00	0.99	1.00	1.00	0.99
LUAD	0.94	0.90	0.96	0.90	0.98	0.93	0.85
LUSC	0.98	0.99	0.98	0.98	1.00	0.97	0.96
PRAD	0.86	0.76	0.91	0.77	0.92	0.82	0.68
STAD	0.90	0.81	0.95	0.85	0.97	0.92	0.77
THCA	0.93	0.90	0.95	0.90	0.98	0.92	0.85
UCEC	0.94	0.89	0.96	0.89	0.99	0.94	0.85
2 group	**Acc**	**Sen**	**Specificity**	**FM**	**AUC**	**Precision**	**CK**

Disease							
BLCA	0.90	0.84	0.93	0.88	0.97	0.90	0.77
BRCA	0.96	0.93	0.97	0.94	0.99	0.95	0.91
KICH	0.99	0.96	1.00	0.98	1.00	1.00	0.97
KIRC	0.98	0.97	0.99	0.98	0.99	0.99	0.97
KIRP	0.96	0.99	0.99	0.99	1.00	0.99	0.98
LUAD	0.95	0.89	0.97	0.91	0.99	0.95	0.86
LUSC	0.99	1.00	0.99	0.99	1.00	0.99	0.99
PRAD	0.89	0.78	0.94	0.80	0.95	0.87	0.73
STAD	0.92	0.83	0.97	0.87	0.98	0.94	0.81
THCA	0.94	0.91	0.96	0.91	0.98	0.93	0.87
UCEC	0.95	0.90	0.97	0.91	1.00	0.95	0.87

Acc:Accuracy, Sen:Sensitivity, FM:F-Measure, AUC:Area Under Curve, CK:Cohen's Kappa

4.4 Comparison Results for all Existing Tools

Comparing all the results obtained is extremely important for bioinformatics tools to be accepted. Because the most important of the possible areas where these tools will be used is the field of health from a life perspective. It is extremely essential to prove that the tools can work correctly and robustly. For this reason, the comparison of the results obtained with maTE, SVM-RCE and SVM-RFE, as well as the miRcorrNet and miRMUTINet tools developed within the scope of this thesis, are presented in Table 4.5. As mentioned before, the outputs of these tools are not the same. While tools other than SVM-RFE perform the classification process through the groups which generated by itself, SVM-RFE does this process at gene level. In this context, to make a fair comparison, group level 2, in other words cluster level 2 results were used for comparison in other tools other than SVM-RFE. In order to include the SVM-RFE tool

in the comparison, it became necessary to determine which gene levels should be used. In this context, the average of the number of genes used in cluster level 2 in all other tools was calculated separately. When these values are examined, it is seen that these numbers vary between 7.45 and 190. Therefore, it was decided to use gene level 8 and gene level 125 results for SVM-RFE.

Table 4.5 Comparison Results Using All 11 Datasets.

Tool Name	#Genes	Acc	Sen	Specificity	AUC	SD
miRcorrNet	141.1	0.96	0.94	0.97	0.98	0.05±0.05
miRMUTINet	78.31	0.96	0.91	0.98	0.99	0.04±0.02
maTE	7.48	0.96	0.94	0.96	0.98	0.034±0.026
SVM-RCE	190.05	0.96	0.94	0.97	0.99	0.06±0.03
SVM-RFE	8	0.84	0.85	0.85	0.91	0.07±0.04
SVM-RFE	125	0.96	0.97	0.95	0.98	0.05±0.03

Column *AUC* is Area Under the Curve, Acc is Accuracy, Sen is Sensitivity, and #Genes is Number of Genes. All the values are averaged over 100 MCCV for the level top 2 groups for maTE and miRcorrNet, while 8 and 125 genes for SVM-RFE and finally for SVM-RCE an average of 190.05 genes from cluster level 2. Standard deviation(SD) values is given for AUC.

4.5 Functional Enrichment Analysis Results

In this section, KEGG pathway enrichment analysis was performed in the light of the results acquired with both miRcorrNet and miRMUTINet. As it is known, all miRNAs and mRNAs are ranked with the RobustRankAggreg package, which is an R package, in both the miRcorrNet tool and the miRMUTINet tool.

When the results produced by the miRcorrNet tool were examined, the number of miRNAs thought to be important in the 11 data sets was too small to be analyzed. For this reason, ranked mRNAs with pValue less than 0.05 were used for miRcorrNet. Accordingly, 85 different enriched KEGG pathways were identified. It is known that these diseases are related to each other. Therefore, the presence of enriched pathways

common to all 11 datasets from the pathways was investigated. In conclusion, no pathway was found to be common in 11 diseases in the results of miRcorrNet. Afterwards, pathways enriched only in each disease itself were investigated. The steps of this process are as follows. To begin with, enriched pathways for all diseases have been identified. Afterwards, enriched pathways in the relevant disease are excluded from this list. Finally, by intersecting these two lists, only pathways enriched in the relevant disease are identified. The results obtained are presented in Table 4.6. Looking at this table, no enriched pathway was found for LUAD and STAD in the analyzes using the results of miRcorrNet. For the other 9 datasets, at least 1 and at most 15 pathways specific to the relevant disease were obtained.

Similarly, when the results produced by the miRMUTINet tool were examined, the number of mRNAs thought to be important in the 11 data sets was too large to be analyzed. For this reason, ranked miRNAs with pValue less than 0.05 were used for miRMUTINet. Accordingly, 157 different enriched KEGG pathways were identified. Using a strategy similar to miRcorrNet, the presence of enriched pathways common to all 11 datasets from the pathways was investigated. The results obtained at this point showed that, unlike miRcorrNet, there are pathways common to all 11 datasets. A total of 55 pathways were found to be common in all 11 disease datasets. By creating a pathway-pathway interaction network from these pathways, it is thought to better elucidate the mechanism of cancer diseases. In this context, thanks to the code written in JAVA language, the similarities of the pathways to each other were calculated with the Kappa score. This formulation shown in Eqs.(4.6). This formulation was used as in this study [63].

$$\kappa = 1 - \frac{1 - p_o}{1 - p_e} \qquad \textbf{(4.6)}$$

A pathway-pathway interaction network was created by using pathway-pathway pairs with a Kappa score above 0.15. The resulting network has been uploaded to Cytoscape [64]. Then, using the cytoHubba plugin, nodes, i.e., pathways, were sorted in the context of MCC and 30 pathways were determined to have MCC values ranging from E14 to E30. The network formed by these pathways is presented in Figure 4.2. As in miRcorrNet, pathways found to be enriched specifically for each disease in

miRMUTINet are presented in Table 4.6. Looking at the results, at least 1 and at most 3 pathways were detected for each disease.

Table 4.6 Specific Enriched Pathways for A Disease Using miRcorrNet and miRMUTINet

Disease	Tool Name	KEGG ID	Pathway Name
BLCA	miRMUTINet	hsa03009	Ribosome biogenesis in eukaryotes
	miRcorrNet	hsa05020	Prion diseases
		hsa04723	Retrograde endocannabinoid signaling
		hsa04919	Thyroid hormone signaling pathway
		hsa04918	Thyroid hormone synthesis
		hsa05200	Pathways in cancer
		hsa04925	Aldosterone synthesis and secretion
		hsa04727	GABAergic synapse
		hsa04911	Insulin secretion
BRCA	miRMUTINet	hsa04720	Long-term potentiation
	miRcorrNet	hsa04144	Endocytosis
KICH	miRMUTINet	hsa04971	Gastric acid secretion
		hsa04514	Cell adhesion molecules (CAMs)
		hsa00240	Pyrimidine metabolism
	miRcorrNet	hsa00260	Glycine, serine and threonine metabolism
		hsa00430	Taurine and hypotaurine metabolism
		hsa00512	Mucin type O-Glycan biosynthesis
		hsa04350	TGF-beta signaling pathway

KICH	miRcorrNet	hsa00590	Arachidonic acid metabolism
		hsa05033	Nicotine addiction
KIRC	miRMUTINet	hsa05014	Amyotrophic lateral sclerosis (ALS)
	miRcorrNet	hsa04742	Taste transduction
KIRP	miRMUTINet	No Specific Pathway Found	
	miRcorrNet	hsa04966	Collecting duct acid secretion
		hsa00534	Glycosaminoglycan biosynthesis - heparan sulfate / heparin
		hsa00061	Fatty acid biosynthesis
		hsa00071	Fatty acid degradation
		hsa01100	Metabolic pathways
		hsa05323	Rheumatoid arthritis
		hsa04974	Protein digestion and absorption
		hsa05110	Vibrio cholerae infection
		hsa04964	Proximal tubule bicarbonate reclamation
		hsa04961	Endocrine and other factor-regulated calcium reabsorption
		hsa05120	Epithelial cell signaling in Helicobacter pylori infection
		hsa00410	beta-Alanine metabolism
		hsa00330	Arginine and proline metabolism
		hsa00051	Fructose and mannose metabolism
		hsa04066	HIF-1 signaling pathway
LUAD	miRMUTINet	hsa04750	Inflammatory mediator regulation of TRP channels
		hsa03460	Fanconi anemia pathway
		hsa05110	Vibrio cholerae infection
	miRcorrNet	No Specific Pathway Found	

LUSC	miRMUTINet	hsa00604	Glycosphingolipid biosynthesis - ganglio series
		hsa04380	Osteoclast differentiation
		hsa04340	Hedgehog signaling pathway
	miRcorrNet	hsa04110	Cell cycle
		hsa03030	DNA replication
		hsa03410	Base excision repair
		hsa03430	Mismatch repair
		hsa00240	Pyrimidine metabolism
		hsa03440	Homologous recombination
		hsa04914	Progesterone-mediated oocyte maturation
		hsa04114	Oocyte meiosis
		hsa01200	Carbon metabolism
		hsa03420	Nucleotide excision repair
		hsa03460	Fanconi anemia pathway
		hsa01059	Biosynthesis of antibiotics
PRAD	miRMUTINet	hsa01040	Biosynthesis of unsaturated fatty acids
		hsa04020	Calcium signaling pathway
		hsa00250	Alanine Aspartate and Glutamate Metabolism
	miRcorrNet	hsa04152	AMPK signaling pathway
		hsa04920	Adipocytokine signaling pathway
		hsa04931	Insulin resistance
STAD	miRMUTINet	No Specific Pathway Found	
	miRcorrNet	No Specific Pathway Found	
THCA	miRMUTINet	hsa00790	Folate biosynthesis
	miRcorrNet	hsa00920	Sulfur metabolism
UCEC	miRMUTINet	hsa05146	Amoebiasis
	miRcorrNet	hsa04510	Focal adhesion

UCEC	miRcorrNet	hsa04014	Ras signaling pathway
		hsa04917	Prolactin signaling pathway
		hsa04711	Circadian rhythm
		hsa04151	PI3K-Akt signaling pathway
		hsa04630	Jak-STAT signaling pathway

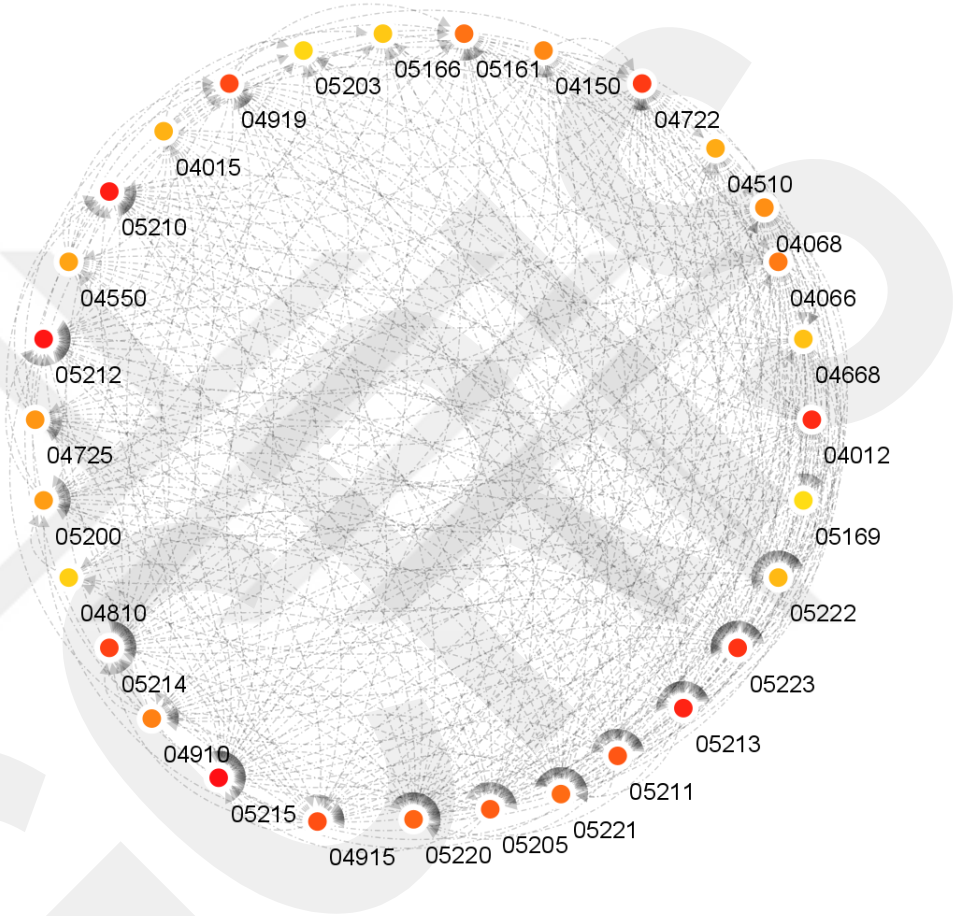

Figure 4.2 Common Pathway-Pathway Interaction Network

4.6 Literature Validation of miRcorrNet Results

In order for a bioinformatics tool to be used as a general standard, the reliability of its results must be proven. In this context, the most reliable method to prove the results obtained is to check whether there is a proof of the existence of the relevant relationship in the literature. The literature has been used to prove the reliability of the results produced with the two tools obtained within the scope of this thesis. In this context, databases in which miRNA - Disease associations are stored in the literature have been

used. These databases are dbDEMC [65], miR2Disease [66], miRCancer [67], PhenomiR [68] and HMDD [69] respectively. Literature validation tables have been created one by one for all diseases. For this validation process, miRNA-Disease pairs produced by running both tools are used. Pairs with a score greater than 1 were used from these pairs. In this way, it is aimed to use the pairs that are the most important for the related disease. The first 3 or 4 pairs with the highest frequency score for all diseases and the most important pair information which no evidence could be found at any database are presented in Table 4.7.

Table 4.7 Validation of miRcorrNet's Results using miRNA – Disease Associations via Existing Databases

BLCA			**BRCA**		
miRNA Name	**Score**	**Evidence**	**miRNA Name**	**Score**	**Evidence**
hsa-miR-21-5p	7.32	dbDEMC, miR2Disease, miRCancer	hsa-miR-21-5p	9.66	dbDEMC, miR2Disease, miRCancer
hsa-miR-22-3p	4.67	miRCancer	hsa-miR-10b-5p	7.98	dbDEMC, miR2Disease, miRCancer
hsa-miR-148b-3p	4.06	dbDEMC, miR2Disease	hsa-miR-200c-3p	5.26	dbDEMC, miR2Disease, miRCancer
KICH			**KIRP**		
hsa-miR-222-3p	9.33	dbDEMC	hsa-miR-21-5p	8.62	dbDEMC, miR2Disease, miRCancer
hsa-miR-221-3p	8.10	dbDEMC, miR2Disease	hsa-miR-10b-5p	4.95	dbDEMC, miR2Disease, miRCancer
hsa-miR-96-5p	7.03	dbDEMC	hsa-miR-589-5p	4.27	dbDEMC
KIRC			**UCEC**		
hsa-miR-28-3p	7.96	dbDEMC, miR2Disease	hsa-miR-151a-5p	2.23	dbDEMC

hsa-miR-21-5p	6.35	dbDEMC, miR2Disease, miRCancer	hsa-miR-200b-3p	2.12	dbDEMC, miRCancer
hsa-miR-106b-3p	6.17	dbDEMC, miR2Disease	hsa-miR-141-3p	2.01	dbDEMC, miRCancer
PRAD			**STAD**		
hsa-miR-143-3p	8.41	dbDEMC, miR2Disease, PhenomiR	hsa-miR-21-5p	9.59	dbDEMC, miR2Disease, miRCancer
has-miR-375	7.72	dbDEMC, miR2Disease	has-miR-148b-3p	3.22	dbDEMC
has-miR-25-3p	6.72	dbDEMC, miR2Disease, miRCancer, PhenomiR	hsa-miR-185-5p	2.39	dbDEMC, miRCancer
hsa-miR-200c-3p	6.51	dbDEMC, miR2Disease, miRCancer, PhenomiR	hsa-miR-200b-3p	2.24	dbDEMC, miR2Disease, miRCancer, PhenomiR
LUAD			**LUSC**		
hsa-miR-30a-3p	6.23	dbDEMC, miR2Disease, miRCancer, HMDD	hsa-miR-146b-3p	3.76	dbDEMC, miR2Disease
hsa-let-7a-5p	6.13	dbDEMC, miR2Disease	hsa-miR-181a-5p	3.74	dbDEMC, miRCancer, PhenomiR,
hsa-miR-22-3p	5.49	dbDEMC	hsa-miR-205-5p	3.44	dbDEMC, miR2Disease
hsa-miR-30c-2-3p	5.36	dbDEMC	hsa-miR-101-3p	3.27	dbDEMC, miR2Disease
THCA					
hsa-miR-152	7.26	dbDEMC			
hsa-miR-30a-5p	6.56	dbDEMC, miRCancer			
hsa-miR-148b-3p	6.50	dbDEMC			

4.7 Literature Validation of miRMUTINet Results

There are 7 different output files obtained after execution of miRMUTINet tool. One of these files is a file that shows the relationship between the miRNAs in the data and the current disease which is under study. The tool ultimately generates the pValue for each miRNA indicating its association with current disease. In this way, miRNAs that thought to be potentially important for the current disease are obtained. All these results were also obtained with mRNAs as well. but since the number of genes with statistical value less than 0.05 was not manageable, miRNAs were used for this process. However, since the number of mRNAs with a pValue of less than 0.05, which is considered statistically significant, is not manageable, miRNAs, not mRNAs, were used for validation. Similar to miRcorrNet, for literature validation of the miRMUTINet's results publicly available miRNA-Disease databases were used. It was investigated whether the obtained miRNA-Disease pairs exist in these databases. Accordingly, approximately 34% of miRNA-Disease pairs, which are considered to be statistically significant, are found in only 1 database, while 23% are in 2 databases, 15% are in 3 databases, 10% are in 4 databases, 6% were found in 5 databases. The number of miRNA-Disease pairs that were not found in the databases, which was shown by miRMUTINet to be related, was determined as 75. The first 3 pairs with the highest pValue for all diseases and the most important pair information which no evidence could be found at any database are presented in Table 4.8.

Table 4.8 Validation of miRMUTINet's Results using miRNA – Disease Associations via Existing Databases

BLCA			**BRCA**		
miRNA Name	**pValue**	**Evidence**	**miRNA Name**	**pValue**	**Evidence**
hsa-miR-16-5p	2.25E-44	dbDEMC, miRCancer, HMDD	hsa-miR-200c-3p	4.22E-98	dbDEMC, miR2Disease, miRCancer, PhenomiR
hsa-miR-1976	1.35E-31	dbDEMC	hsa-miR-378a-5p	3.74E-95	dbDEMC

hsa-miR-141b-5p	1.35E-31	dbDEMC, miR2Disease	hsa-miR-200b-3p	1.64E-92	dbDEMC, miR2Disease, miRCancer, PhenomiR
hsa-miR-24-2-5p	-	*No evidence*	*hsa-miR-500a-3p*	-	*No evidence*
KICH			**KIRP**		
hsa-miR-508-3p	1.9E-112	dbDEMC, miRCancer	hsa-miR-125a-5p	1.6E-139	dbDEMC
hsa-miR-26b-3p	2.61E-98	dbDEMC	hsa-miR-338-3p	1.27E-68	dbDEMC
hsa-miR-222-3p	3.94E-77	dbDEMC	hsa-miR-139-5p	1.08E-66	dbDEMC
hsa-miR-874	*6.8E-100*	*No Evidence*	*hsa-miR-500b*	*7.62E-65*	*No Evidence*
KIRC			**UCEC**		
hsa-miR-200b-3p	7.79E-82	dbDEMC, miR2Disease	hsa-miR-106b-3p	1.99E-53	dbDEMC, miRCancer
hsa-miR-200c-3p	6.18E-66	dbDEMC, miR2Disease, miRCancer, PhenomiR	hsa-miR-140-3p	1.35E-52	dbDEMC
hsa-miR-21-5p	2.69E-63	dbDEMC, miR2Disease, miRCancer	hsa-miR-200a-3p	5.28E-52	dbDEMC, miRCancer
hsa-miR-203a	*2.82E-79*	*No Evidence*	*hsa-miR-128*	*6.58E-51*	*No Evidence*
PRAD			**STAD**		
hsa-miR-25-3p	2.57E-91	dbDEMC, miR2Disease, miRCancer, PhenomiR	hsa-miR-155-5p	2.75E-43	miRCancer, HMDD
hsa-miR-93-5p	2.95E-87	dbDEMC, miRCancer, PhenomiR	hsa-miR-330-5p	1.1E-42	miRCancer
hsa-miR-182-5p	8.25E-73	dbDEMC, miR2Disease, miRCancer, PhenomiR	hsa-miR-30c-2-3p	3.16E-42	dbDEMC

hsa-miR-769-5p	*9.01E-13*	*No Evidence*	*hsa-miR-942*	*4.84E-41*	*No Evidence*
LUAD			**LUSC**		
hsa-miR-30a-3p	3.09E-83	dbDEMC, miR2Disease, miRCancer, PhenomiR, HMDD	hsa-miR-181a-5p	4.83E-58	dbDEMC, miRCancer, PhenomiR,
hsa-miR-143-3p	2.14E-78	dbDEMC, miR2Disease, PhenomiR	hsa-miR-126-5p	2.79E-57	dbDEMC, miR2Disease, miRCancer, PhenomiR, HMDD
hsa-miR-126-5p	6.26E-77	miR2Disease, PhenomiR, HMDD	hsa-miR-140-3p	5.9E-55	dbDEMC, miR2Disease, miRCancer, PhenomiR, HMDD
THCA					
hsa-miR-221-3p	6.2E-135	dbDEMC, miR2Disease, miRCancer, PhenomiR, HMDD			
hsa-miR-222-3p	1.5E-125	dbDEMC, miR2Disease, miRCancer, PhenomiR, HMDD			
hsa-miR-34a-5p	1.92E-77	dbDEMC, miR2Disease, HMDD			

4.8 External Data Validation of miRcorrNet Results

For validation processes, in addition to the literature validation, the performance of the model should be evaluated using an external dataset. are transactions. For this purpose, RNA-seq data with access code GSE40419 was used [70]. This data set consists of a total of 164 samples, including 87 lung adenocarcinoma tissue and 77 normal tissue data. A separate KNIME workflow has been developed for test operations on external data. This KNIME workflow takes 4 separate files as input. The first file is the mRNA file used to train the test workflow which in this case this data is LUAD data. The second file is the RNA-seq data file to be used for testing which in this case is GSE40419 (i.e., LUAD_E). In addition to these two files, a certain number of randomly given genes and the same number of genes found to be potentially important by the miRcorrNet tool are given. In this way, the results of random genes and important genes

found by miRcorrNet could be directly compared. The numbers of genes given in this context are 1, 5, 30 and 50, respectively. The results obtained are presented in Table 4.9. Looking at the results in Table 4.9, when only 1 gene was used, the result of the miRcorrNet tool produced a 45% more successful result compared to the randomly used gene, while this situation was 28% in 5 genes, 6% in 30 genes and 8% in 50 genes. As can be understood from the values specified here, the results produced by the miRcorrNet tool are also robust on external data.

Table 4.9 External Data Validation Results of miRcorrNet

Experiments	Sensitivity	Specificity	Accuracy
LUAD (train) test on LUAD_E random 1	0.53	0.61	0.56
LUAD (train) test on LUAD_E random 5	0.73	0.76	0.74
LUAD (train) test on LUAD_E random 30	0.87	0.94	0.91
LUAD (train) test on LUAD_E random 50	0.88	0.93	0.90
LUAD (train) test on LUAD_E top 1	0.86	0.75	0.81
LUAD (train) test on LUAD_E top 5	0.97	0.92	0.95
LUAD (train) test on LUAD_E top 30	0.98	0.97	0.97
LUAD (train) test on LUAD_E top 50	0.99	0.97	0.98

4.9 External Data Validation of miRMUTINet Results

In addition to the literature validation, external data performance validation was performed for the miRMUTINet tool. For this process, just like in miRcorrNet, RNA-seq data with access code GSE40419 was used as external data (i.e., LUSC_E). LUSC mRNA data was used for the train process of the test workflow. Validation results were obtained using the KNIME test workflow and the previously mentioned datasets, and these results are presented in Table 4.10. When the results in Table 4.10 were examined, the accuracy was 51% when the classification was performed using random 1 gene. For

comparison purposes, this process was repeated using the gene that miRMUTINet found most important and accuracy increased by 71% to 87%. When the increases in Accuracy are examined, it has been shown that the use of the miRMUTINet tool leads to higher performance. Looking at the results, accuracy reached 88% with a 60% increase when using 5 genes, accuracy reached 93% with a 22% increase when using 30 genes, and finally, accuracy reached 95% with a 10% increase when using 50 genes. Looking at these values, it is seen that the results of miRMUTINet are reliable and noteworthy.

Table 4.10 External Data Validation Results of miRMUTINet

Experiments	**Sensitivity**	**Specificity**	**Accuracy**
LUSC (train) test on LUSC_E random 1	0.43	0.58	0.51
LUSC (train) test on LUSC_E random 5	0.46	0.61	0.55
LUSC (train) test on LUSC_E random 30	0.57	0.91	0.76
LUSC (train) test on LUSC_E random 50	0.76	0.94	0.86
LUSC (train) test on LUSC_E top 1	0.84	0.88	0.87
LUSC (train) test on LUSC_E top 5	0.94	0.81	0.88
LUSC (train) test on LUSC_E top 30	0.94	0.92	0.93
LUSC (train) test on LUSC_E top 50	0.94	0.97	0.95

Chapter 5

DISCUSSIONS

The aim of this thesis is to develop a machine learning model that will separate the samples belonging to the case and control groups, as mentioned earlier. For this purpose, publicly available miRNA and mRNA data which obtained from cancer patients were used. It is very difficult to develop a machine learning model in such diseases which has a complex molecular mechanism such as cancer. In order to overcome this shortcoming, it seems essential to integrate and use different types of -omic data. In this way, it is planned that the machine learning model to be developed will be effective and that the results will be robust and reliable.

These planned applications were implemented, and a machine learning model was obtained in this direction. Results were obtained with the model developed for miRcorrNet. These results were examined within the scope of AUC, and it was seen that the values varied between 0.93 and 1.00. In addition, it has been observed that successful results have been obtained in the test processes performed on different data sets that have never been used before. Different information levels were compared for the test processes of miRcorrNet. These levels are 1, 5, 30 and 50 genes, respectively. As shown in Table 4.8.1, the workflow was run using random genes and the genes that miRcorrNet found to be important, and the effectiveness of the model was tried to be demonstrated. When the performance results of a randomly selected gene against the most important gene found by miRcorrNet were compared, it was determined that miRcorrNet was 45% more successful in terms of accuracy. Similarly, this ratio was 28% when 5 genes were used, 6% when 30 genes were used, and 8% when 50 genes were used. It has been observed that this success has also been achieved for the other developed bioinformatics tool, miRMUTINet. miRMUTINet's results were examined within the scope of AUC, and it was seen that the values varied between 0.92 and 1.00. The results obtained using external data for testing purposes and these results were also

considered noteworthy. The same strategy as miRcorrNet was used in the testing step. When the performance results of a randomly selected gene against the most important gene found by miRMUTINet were compared, it was determined that miRMUTINet was 71% more successful in terms of accuracy. Similarly, this ratio was 60% when 5 genes were used, 22% when 30 genes were used, and 10% when 50 genes were used.

The results obtained with miRcorrNet were obtained for 11 different cancer datasets. The latest version of miRBase, 22.1, contains approximately 2500 miRNAs [71]. miRcorrNet has prioritized very few miRNAs for all these different diseases (minimum: 13 and maximum:92 miRNAs). It was also examined whether there were common miRNAs prioritized for these different diseases. In this context, it was seen that 11 miRNAs were prioritized in 6 different cancer types. These miRNAs are presented in Figure 5.1.

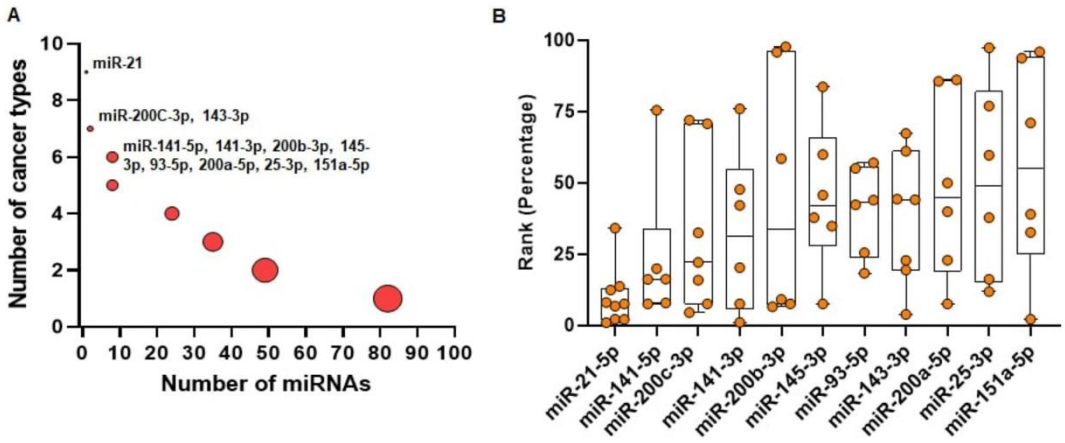

Figure 5.1 miRNA Analysis for miRcorrNet.(A) Eleven miRNAs the potentially regulate 6 or more cancer types, are highlighted. (B) Ranks of these 11 miRNAs in individual cancer types are denoted by dots. These miRNAs are sorted based on their median rank.

Some of these miRNAs have been prioritized by miRcorrNet in a large number of diseases as seen in Figure 5.1.A. Not only was it prioritized, but these miRNAs were also seen to be prioritized on top as seen in Figure 5.1.B . hsa-miR-21-5p has been associated with 9 of 11 cancer types. In addition, hsa-miR-200c and hsa-miR-143-3p miRNAs have been associated with 7 cancers. One of these miRNAs, hsa-miR-21, is a very important onco-miRNA that represses genes that act as tumor suppressors and is

also directly related to multiple cancers [72,73]. In addition to all these, it is known that hsa-miR-21 has a diagnostic role for cancer types. The effect of resistance against the developed drugs has also been the subject of literature studies [74–76].

Any information that can be obtained in diseases where the molecular mechanism and functionality is complex is very important. Hence, it is essential to obtain the big network of the disease, in other words the big picture of the disease, and to reveal the existing interactions i.e., relationships among them. For this purpose, a pathway - pathway interaction network has been produced. When creating this network, pathways from all 11 disease datasets were used. Because its presence in all datasets may prove that it is indeed associated with the disease. It is also necessary to determine the pathways that are more important than the others among the pathways that are each node in the generated network. For this purpose, the Cytoscape plugin cytoHubba was used. As a result, Matthews Correlation Coefficient values were generated for each node and pathways were ranked. A part of the results obtained is also presented in Table 5.1. When the scores produced were examined, it was observed that the first 30 pathways had very high MCC scores, which is sufficient to see that these pathways are important for understanding the mechanism of cancer.

Table 5.1 Ranked Common Pathways Using Pathway – Pathway Interaction Network

KEGG ID	Pathway Name	MCC	Degree
05223	Non-small cell lung cancer	9.22E28	27
05221	Acute myeloid leukemia	9.22E28	26
05220	Chronic myeloid leukemia	9.22E28	30
05215	Prostate cancer	9.22E28	31
05214	Glioma	9.22E28	30
05213	Endometrial cancer	9.22E28	29
05212	Pancreatic cancer	9.22E28	30
05211	Renal cell carcinoma	9.22E28	26

The classes of these pathways in the BRITE hierarchy were also examined, and it was desired to have an idea about their functionality. In this context, it has been observed that 3 of the 9 pathways enriched within the scope of BLCA are related to the nervous system, while 5 of them are related to the endocrine system. One of the 2 pathways found for BRCA was related to the nervous system, while one of them was related to the cellular process. It is noteworthy that 5 of the 9 pathways found for KICH are metabolic pathways. In addition, pathways related to the Digestive system were also found. It was determined that one of the 2 pathways obtained for KIRC was related to the neurodegenerative system, while the other pathway was associated with the sensory system. The 15 pathways found for KIRP revealed a more complex network of relationships. In this context, it has been observed that there are pathways associated with both the excretory system, the digestive system and the Immune disease pathway. In addition, it has been observed that the relationship with metabolic pathways, especially lipid metabolism, is important with respect to this disease. Another interesting point is that *hsa05510* (*Vibrio cholerae infection*) and *hsa05120* (*Epithelial cell signaling in Helicobacter pylori infection),* which are known to be associated with Infectious disease, may be important for this disease. In the 3 pathways found for LUAD, pathways related to both the sensory system and genetic information were found. For LUSC, a remarkable 6 of the 15 pathways found are pathways related to Genetic information processing. This shows that genetic factors could be a serious reason for this disease. In addition, 3 of the remaining pathways are different metabolic pathways while 3 are cellular processes related pathways. A total of 6 pathways were found for PRAD. While 2 of these pathways are different metabolic pathways, 2 of them were found to be related to the Endocrine system. Interestingly, no enriched pathway for STAD was found in either miRcorrNet or miRMUTINet. It has been observed that the 2 pathways found for THCA are also different metabolic pathways. Finally, a total of 7 different pathways were found for UCEC. There was not enough information for 2 of these pathways. The remaining 5 pathways were found to belong to different classes from each other which indicates that the bigger picture of UCEC is more complex among others.

There are several reasons why any developed bioinformatic tool is accepted as a general standard application. The most important of these are the satisfactory performance results, the satisfactory success in the procedures performed for the test purposes using external data, the biologically relevant and explanatory results, and lastly is to prove the reliability of the results obtained by providing literature validation. Because of the last aforementioned reason, literature validation processes were carried out for both tools developed in this context. For this process, databases used in the literature and in which the relationships between diseases and miRNAs are kept were used. miRNAs were used for validation because the number of important genes obtained for cancer types averages just over 1000. This showed that it is not very feasible to work on genes. However, not all miRNAs obtained as output were also used. For this filtering process, the score parameter produced in the tool was considered and miRNAs with a score value greater than 1.00 were preferred for validation processes.

A summary of the results of the literature validation is presented in Table 5.2. Looking at the table, there are 147 miRNA-Disease associations used for validation by filtering from the results produced by the miRcorrNet tool. 56 of these pairs were found in only 1 database. Likewise, while 57 of them were found in 2 databases, 24 of them were found in 3 databases and 7 of these pairs were found in 4 databases. By filtering the results produced by the miRMUTINet tool, it was determined that there were 682 miRNA-Disease associations. Of these pairs, roughly 34% were found in only 1 database, while 23% were found in 2 databases, 15% in 3 databases, 10% in 4 databases, and 6% in 5 databases simultaneously. It has been seen that the results found in this context are in great agreement with the literature and the results are robust and biologically relatable.

In this thesis we aimed 3 different things. First, we wanted to find important biomarkers as miRNA and as mRNA to fight cancer. Considering the produced results, we believe that we acquired successful results in terms of finding biomarkers. For example, it has been observed that the accuracy metric obtained by classification with the most important gene found by miRcorrNet is 45% better than the accuracy obtained by using a random gene. In comparison at the same gene level, miRMUTINet achieved 71% better accuracy. This shows that both of our tools found important biomarkers. Second, we wanted to integrate multi-omic data to explore important miRNA - mRNA

groups. We acquired all important groups both using statistics-based metric and correlation-based metric. The produced groups were subjected to the ranking procedure and the groups with the best classification performance were determined. When the results of machine learning models are examined, it is thought that groups may also be biologically important for related diseases. Finally, we wanted to enlighten the cancer mechanisms as we much as we can. By using multi-omic data we tried to find enriched pathways and their interactions. For this purpose, we generated pathway - pathway interaction network and we saw that we found 30 important pathways in terms of MCC value, and we found 403 edges i.e., interactions between them. We believe that this inter-relationship is noteworthy.

Table 5.2 Summary of miRNA-Disease Relations Found in Existing Databases

Disease	**Number of miRNA-Disease Associations**	**Number of databases containing the specific miRNA - Disease Association**				
		1	**2**	**3**	**4**	**5**
miRcorrNet						
BLCA	14	5	7	1	-	-
BRCA	7	-	2	4	1	-
KICH	11	9	2	-	-	-
KIRC	13	8	3	2	-	-
KIRP	17	10	5	2	-	-
LUAD	14	3	5	4	2	-
LUSC	22	4	15	2	1	-
PRAD	7	1	3	3	-	-
STAD	9	2	4	3	-	-
THCA	12	4	4	2	2	-
UCEC	21	10	7	1	1	-
miRMUTINet						
BLCA	62	21	17	9	6	2

BRCA	51	4	15	19	11	-
KICH	61	34	15	-	-	-
KIRC	46	27	9	5	-	-
KIRP	87	44	19	4	-	-
LUAD	91	11	26	31	15	8
LUSC	54	2	6	10	15	20
PRAD	53	9	11	14	13	4
STAD	35	8	14	6	4	2
THCA	55	28	9	8	2	4
UCEC	87	46	20	-	-	-

Chapter 6

Conclusions and Future Studies

6.1 Conclusions

Bioinformatics discipline emerges as a very important field with its multidisciplinary working structure. Especially when we look at the structure of the data produced today, the interaction of bioinformatics with fields such as data mining, statistics and machine learning has become much more critical. The combined use of all these interacting areas provides a wide range of studies, from finding potential biomarkers for diseases to affected pathways Thanks to such studies, data can be collected to accelerate drug design and development processes. As it can be understood from this statement, thanks to these drugs that will be developed against diseases, treatment opportunities will increase, and hopefully the number of deaths will be reduced. By using this know-how obtained as a result of these studies, it will be possible to develop not only collective treatment methods, but also personalized treatment methods. In this way, it will be possible to reduce the number of approximately 10 million deaths due to cancer.

Detection of potential biomarkers for cancer types is important for aforementioned reasons. In this context, it is a necessity to use more data in order to produce more robust results. Therefore, it is considered that using different -omic data together will be more effective both for the machine learning model to be developed and within the scope of the results to be obtained. For this purpose, within the scope of this thesis, two bioinformatic tools (miRcorrNet and miRMUTINet) have been developed for the detection of cancer patients by integrating both miRNA and mRNA data. In both developed tools, primarily differentially expressed mRNAs and miRNAs were detected. Subsequently, mRNA-miRNA groups were formed. The mRNAs in these groups are the mRNAs that are thought to be related to the miRNA in their group

in the context of the metric used. Afterwards, the performance of these groups to distinguish two classes (control and case) from each other was determined by a ranking procedure. At the end of this procedure, it is known that the generated groups definitely distinguish these two classes.

Seven different files are created as a result of running these implemented KNIME workflows. In this way, all outputs provide satisfactory information for the purpose. The obtained results were analyzed both at the mRNA level, at the miRNA level and at the integrated mRNA-miRNA group level. In this context, mRNAs, miRNAs and mRNA-miRNA groups considered to be potentially important were identified. All results were validated using two methods. The first method is the validation of the miRNA- Disease relationships produced by the tools. In this context, the miRNA - Disease pairs and the entries in the databases where the miRNA - Disease relations are stored were compared. When the results were examined, it was observed that both the results of miRcorrNet and the results of miRMUTINet had a very high validity in the context of the literature. The second method is to run a data set that has never been used before on the developed test tool. In this way, it is expected to prove that the developed tool does not work with chance. As explained in the results section, the results were obtained and compared by using 4 different levels (1,5,30 and 50 genes). It has been proven that the results produced by the tools are not obtained randomly.

It is also extremely important that the results are biologically consistent. In this context, enriched pathways using miRNA and mRNA activities were identified using various online tools. In order to understand the molecular mechanism of cancer, which has a complex structure, enriched pathways in 11 different datasets were examined. With miRcorrNet, all datasets did not have pathways that were commonly enriched, whereas with miRMUTINet, 55 pathways were found that were commonly enriched in all datasets. All of those commonly enriched pathways were ranked according to their own MCC scores, and it was observed that the first 30 pathways had very close MCC scores. This gives an idea of the big picture that needs to be looked at for cancer.

6.2 Societal Impact and Contribution to Global Sustainability

In this thesis, it is planned to contribute to public health in the medical sense by using bioinformatic methods. Thanks to the methods developed in this context, it was thought to contribute to the drug design processes by trying to obtain biomarkers of cancer types. Following the proof of the information in this thesis by the clinicians, it was envisaged that the drugs could be produced both for the community and designed for rare cases.

The effects of miRNAs on the functioning of the body are known. In addition to this situation, it has been observed that miRNAs are very effective in the formation of various diseases. There are also mRNAs that play an important role in functioning in our body. It was planned to reveal the miRNA - mRNA relationships that affect the formation and progression of diseases by the use of the above-mentioned miRNA and mRNA, in other words, multiple -omics data, and this aim was achieved as is planned. In this context, two different bioinformatics tools were produced. These tools also include the Random Forest algorithm, which is one of the state-of-the-art machine learning techniques. In this way, it has been possible to reveal relationships that are almost impossible to obtain with traditional methods. In addition, biomarkers that best distinguish patients from normal individuals were ranked by a ranking procedure, and miRNA-mRNA relationships, which are considered to be potentially the most important in the context of the disease, were revealed. In addition to miRNA - mRNA interaction pairs, enriched pathways based on these pairs have also been identified. The results obtained have been extensively compared both with existing miRNA-Disease databases and with other existing tools. As a result of this thesis, it is aimed to contribute to the society by revealing the relationships that have both direct and indirect effects on diseases.

6.3 Future Prospects

A detailed study was carried out within the scope of this thesis. However, there are points that need improvement. The first of these points is related to the groups used in the classification process. In both miRcorrNet and miRMUTINet, groups are used singularly. In other words, performance evaluation was not carried out by determining

the relations between the groups and using them in an ensemble manner. It is thought that with the new method to be developed by adopting this approach, more robust and higher success in terms of performance will be achieved.

The second point is that although the realized machine learning model produces successful results, considering the structure of the data (lower number of sample vs higher number of features), it is desired to obtain the results to be obtained with the deep learning approach. As a result, by comparing the traditional state-of-the-art machine learning techniques with the deep learning method, which is a relatively more innovative paradigm, it will be determined which paradigm achieves more successful results.

6.4 Availability

All works carried out within the scope of this thesis are kept in 3 different GitHub repository. All materials for miRcorrNet bioinformatics tool can be accessed from https://github.com/gokhangoy/miRcorrNet web address, and all materials for miRMUTINet bioinformatics tool can be accessed from https://github.com/gokhangoy /miRModuleNet web address. The materials in these two repositories are respectively: used data(/Data), produced figures(/Figures), implemented KNIME workflows (/KNIME_Workflows), obtained results (/Results), used visualization script(/ Scripts) and all outputs obtained for validation of results (/Validation). Apart from these two repositories, the code used to generate the pathway - pathway interaction network can be accessed from https://github.com/gokhangoy/pathwayClusteringTool repository.

BIBLIOGRAPHY

[1] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F., "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries", CA: A Cancer Journal For Clinicians, (2020): .

[2] Bartel, D. P., "MicroRNAs: Genomics, Biogenesis, Mechanism, and Function", Cell, 116 (2), 281–297 (2004).

[3] Cai, Y., Yu, X., Hu, S., and Yu, J., "A Brief Review on the Mechanisms of miRNA Regulation", Genomics, Proteomics & Bioinformatics, 7, 147–154 (2009).

[4] Cheng, A. M., "Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis", Nucleic Acids Research, 33, 1290–1297 (2005).

[5] Keller, A., Leidinger, P., Bauer, A., ElSharawy, A., Haas, J., Backes, C., Wendschlag, A., Giese, N., Tjaden, C., Ott, K., Werner, J., Hackert, T., Ruprecht, K., Huwer, H., Huebers, J., Jacobs, G., Rosenstiel, P., Dommisch, H., Schaefer, A., Müller-Quernheim, J., Wullich, B., Keck, B., Graf, N., Reichrath, J., Vogel, B., Nebel, A., Jager, S. U., Staehler, P., Amarantos, I., Boisguerin, V., Staehler, C., Beier, M., Scheffler, M., Büchler, M. W., Wischhusen, J., Haeusler, S. F. M., Dietl, J., Hofmann, S., Lenhof, H.-P., Schreiber, S., Katus, H. A., Rottbauer, W., Meder, B., Hoheisel, J. D., Franke, A., Meese, E., "Toward the blood-borne miRNome of human diseases", Nature Methods, 8, 841–843 (2011).

[6] Tüfekci, K. U., Öner, M. G., Meuwissen, R. L. J., Genç, Ş., "The Role of MicroRNAs in Human Diseases", MiRNomics: MicroRNA Biology and Computational Analysis, Humana Press, Totowa, NJ, 33–50 (2014).

[7] Calin, G. A. and Croce, C. M., "MicroRNA signatures in human cancers", Nature Reviews Cancer, 6, 857–866 (2006).

[8] Berindan-Neagoe, I., Monroig, P., Pasculli, B., and Calin, G. A., "MicroRNAome genome: a treasure for cancer diagnosis and therapy", CA: A Cancer Journal For Clinicians, 64, 311–336 (2014).

[9] Phuah, N. H, Nagoor, N. H., "Regulation of MicroRNAs by Natural Agents: New Strategies in Cancer Therapies", BioMed Research International, (2014).

[10] Cohen, J., "Bioinformatics—an introduction for computer scientists", ACM Computing Surveys, 36, 122–158 (2004).

[11] Yousef, M., Abdallah, L., Allmer, J., "MaTE: discovering expressed interactions between microRNAs and their targets", Bioinformatics, 35, 4020–4028 (2019).

[12] Pencheva, N., Tavazoie, S. F., "Control of Metastatic Progression by microRNA Regulatory Networks", Nature Cell Biology, 15, 546–554 (2013).

[13] Lai, E. C., "Micro RNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation", Nature Genetics, 30, 363–364 (2002).

[14] Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., Kellis, M., "Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals", Nature, 434, 338–345 (2005).

[15] Riffo-Campos, Á. L., Riquelme, I., Brebi-Mieville, P., "Tools for Sequence-Based miRNA Target Prediction: What to Choose?", International Journal Of Molecular Sciences, 17, (2016).

[16] Ivey, K. N., Srivastava, D., "MicroRNAs as Developmental Regulators", Cold Spring Harbor Perspectives In Biology, 7, (2015).

[17] Yousef, M., Goy, G., Mitra, R., Eischen, C. M., Jabeer, A., Bakir-Gungor, B., "MiRcorrNet: machine learning-based integration of miRNA and mRNA expression profiles, combined with feature grouping and ranking", PeerJ, 9, (2021).

[18] Vasudevan, S., Tong, Y., Steitz, J. A., "Switching from Repression to Activation: MicroRNAs Can Up-Regulate Translation", Science, 318, 1931–1934 (2007).

[19] Feng, Y., Xing, Y., Liu, Z., Yang, G., Niu, X., Gao, D., "Integrated analysis of microRNA and mRNA expression profiles in rats with selenium deficiency and identification of associated miRNA-mRNA network", Scientific Reports, 8, 6601 (2018).

[20] Liu, Y., Zhang, J., Xu, Q., Kang, X., Wang, K., Wu, K., Fang, M., "Integrated miRNA-mRNA analysis reveals regulatory pathways underlying the curly fleece trait in Chinese tan sheep", BMC Genomics, 19, 360 (2018).

[21] Li, L., Peng, M., Xue, W., Fan, Z., Wang, T., Lian, J., Zhai, Y., Lian, W., Qin, D., Zhao, J., "Integrated analysis of dysregulated long non-coding RNAs/microRNAs/mRNAs in metastasis of lung adenocarcinoma", Journal Of Translational Medicine, 16, 372 (2018).

[22] Yao, Y., Jiang, C., Wang, F., Yan, H., Long, D., Zhao, J., Wang, J., Zhang, C., Li, Y., Tian, X., Wang, Q. K., Wu, G., Zhang, Z., "Integrative Analysis of miRNA and mRNA Expression Profiles Associated With Human Atrial Aging", Frontiers In Physiology, 10, (2019).

[23] Yang, L., Li, L., Ma, J., Yang, S., Zou, C., Yu, X., "MiRNA and mRNA Integration Network Construction Reveals Novel Key Regulators in Left-Sided and Right-Sided Colon Adenocarcinoma", BioMed Research International, 2019, 1–9 (2019).

[24] Huang, J. C., Morris, Q. D., Frey, B. J., "Bayesian Inference of MicroRNA Targets from Sequence and Expression Data", Journal Of Computational Biology, 14, 550–563 (2007).

[25] Huang, J. C., Frey, B. J., Morris, Q. D., "COMPARING SEQUENCE AND EXPRESSION FOR PREDICTING microRNA TARGETS USING GenMiR3", Proceedings of the Pacific Symposium, Kohala Coast, Hawaii, USA, (2007).

[26] Stingo, F. C., Chen, Y. A., Vannucci, M., Barrier, M., Mirkes, P. E., "A BAYESIAN GRAPHICAL MODELING APPROACH TO MICRORNA REGULATORY NETWORK INFERENCE", The Annals Of Applied Statistics, 4, 2024–2048 (2010).

[27] Le, H.-S. and Bar-Joseph, Z., "Integrating sequence, expression and interaction data to determine condition-specific miRNA regulation", Bioinformatics, 29, i89–i97 (2013).

[28] Luo, J., Pan, C., Xiang, G., Yin, Y., "A Novel Cluster-Based Computational Method to Identify miRNA Regulatory Modules", IEEE/ACM Transactions On Computational Biology And Bioinformatics, 16, 681–687 (2019).

[29] Friedman, N., Linial, M., Nachman, I., Pe'Er, D., "Using Bayesian Networks to Analyze Expression Data", 20, (2014) .

[30] Liu, B., Li, J., Tsykin, A., Liu, L., Gaur, A. B., Goodall, G. J., "Exploring complex miRNA-mRNA interactions with Bayesian networks by splitting-averaging strategy", BMC Bioinformatics, 10, 408 (2009).

[31] Jayaswal, V., Lutherborrow, M., Ma, D. D., Yang, Y. H., "Identification of microRNA-mRNA modules using microarray data", BMC Genomics, 12, 138 (2011).

[32] Hecker, N., Stephan, C., Mollenkopf, H.-J., Jung, K., Preissner, R., Meyer, H.-A., "A New Algorithm for Integrated Analysis of miRNA-mRNA Interactions Based on Individual Classification Reveals Insights into Bladder Cancer", PLoS ONE, 8, e64543 (2013).

[33] Joung, J.-G., Hwang, K.-B., Nam, J.-W., Kim, S.-J., Zhang, B.-T., "Discovery of microRNA mRNA modules via population-based probabilistic learning", Bioinformatics, 23, 1141–1147 (2007).

[34] Baluja, S., "Population-Based Incremental Learning:", 41 (1994).

[35] Larrañaga, P., Lozano, J. A., "Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation", Springer Science & Business Media, 424 (2001).

[36] Zhang, B.-T., "A Unified Bayesian Framework for Evolutionary Learning and Optimization", Advances in Evolutionary Computing, Springer Berlin Heidelberg, Berlin, Heidelberg, 393–412 (2003).

[37] Zhang, S., Li, Q., Liu, J., Zhou, X. J., "A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules", Bioinformatics, 27, i401–i409 (2011).

[38] Tran, D. H., Satou, K., Ho, T. B., "Finding microRNA regulatory modules in human genome using rule induction", BMC Bioinformatics, 9, 5 (2008).

[39] Lavracˇ, N., Kavsˇek, B., Flach, P., Todorovski, L., "Subgroup Discovery with CN2-SD", Journal of Machine Learning Research, 5, 183–188 (2004).

[40] Paul, S., Lakatos, P., Hartmann, A., Schneider-Stock, R., Vera, J., "Identification of miRNA-mRNA Modules in Colorectal Cancer Using Rough Hypercuboid Based Supervised Clustering", Scientific Reports, 7, 42809 (2017).

[41] Yousef, M., Jung, S., Showe, L. C., Showe, M. K., "Recursive Cluster Elimination (RCE) for classification and feature selection from gene expression data", BMC Bioinformatics, 8, 144 (2007).

[42] Yousef, M., Bakir-Gungor, B., Jabeer, A., Goy, G., Qureshi, R., C. Showe, L., "Recursive Cluster Elimination based Rank Function (SVM-RCE-R) implemented in KNIME", F1000Research, 9, (2021).

[43] "GDC", https://portal.gdc.cancer.gov/ (13.02.2021).

[44] "The Cancer Genome Atlas Program - National Cancer Institute", https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga (13.02.2021).

[45] Mitra, R., Adams, C. M., Jiang, W., Greenawalt, E., Eischen, C. M., "Pan-cancer analysis reveals cooperativity of both strands of microRNA that regulate tumorigenesis and patient survival", Nature Communications, 11, 968 (2020).

[46] "Comprehensive Molecular Characterization of Urothelial Bladder Carcinoma", Nature, 507, 315–322 (2014).

[47] Nederlof, I., De Bortoli, D., Bareche, Y., Nguyen, B., de Maaker, M., Hooijer, G. K. J., Buisseret, L., Kok, M., Smid, M., Van den Eynden, G. G. G. M., Brinkman, A. B., Hudecek, J., Koster, J., Sotiriou, C., Larsimont, D., Martens, J. W. M., van de Vijver, M. J., Horlings, H. M., Salgado, R., Biganzoli, E., Desmedt, C., "Comprehensive evaluation of methods to assess overall and cell-specific immune infiltrates in breast cancer", Breast Cancer Research , 21, (2019).

[48] Davis, C. F., Ricketts, C., Wang, M., Yang, L., Cherniack, A. D., Shen, H., Buhay, C., Kang, H., Kim, S. C., Fahey, C. C., Hacker, K. E., Bhanot, G., Gordenin, D. A., Chu, A., Gunaratne, P. H., Biehl, M., Seth, S., Kaipparettu, B. A., Bristow, C. A., Donehower, L. A., Wallen, E. M., Smith, A. B., Tickoo, S. K., Tamboli, P., Reuter, V., Schmidt, L. S., Hsieh, J. J., Choueiri, T. K., Hakimi, A. A., Chin, L., Meyerson, M., Kucherlapati, R., Park, W.-Y., Robertson, A. G., Laird, P. W., Henske, E. P., Kwiatkowski, D. J., Park, P. J., Morgan, M., Shuch, B., Muzny, D., Wheeler, D. A., Linehan, W. M., Gibbs, R. A., Rathmell, W. K., Creighton, C. J.,

"The somatic genomic landscape of chromophobe renal cell carcinoma", Cancer Cell, 26, 319–330 (2014).

[49] Lee, B. H., "Commentary on: “Comprehensive molecular characterization of papillary renal-cell carcinoma.” Cancer Genome Atlas Research Network.", Urologic Oncology: Seminars And Original Investigations, 35, 578–579 (2017).

[50] "COMPREHENSIVE MOLECULAR CHARACTERIZATION OF CLEAR CELL RENAL CELL CARCINOMA", Nature, 499, 43–49 (2013).

[51] "Comprehensive molecular profiling of lung adenocarcinoma", Nature, 511, 543–550 (2014).

[52] "Comprehensive genomic characterization of squamous cell lung cancers", Nature, 489, 519–525 (2012).

[53] "The molecular taxonomy of primary prostate cancer", Cell, 163, 1011–1025 (2015).

[54] Bass, A. J., Thorsson, V., Shmulevich, I., Reynolds, S. M., Miller, M., Bernard, B., Hinoue, T., Laird, P. W., Curtis, C., Shen, H., Weisenberger, D. J., Schultz, N., Shen, R., Weinhold, N., Kelsen, D. P., Bowlby, R., Chu, A., Kasaian, K., Mungall, A. J., Robertson, A. G., Sipahimalani, P., Cherniack, A., Getz, G., Liu, Y., Noble, M. S., Pedamallu, C., Sougnez, C., Taylor-Weiner, A., Akbani, R., Lee, J.-S., Liu, W., Mills, G. B., Yang, D., Zhang, W., Pantazi, A., Parfenov, M., Gulley, M., Piazuelo, M. B., Schneider, B. G., Kim, J., Boussioutas, A., Sheth, M., Demchok, J. A., Rabkin, C. S., Willis, J. E., Ng, S., Garman, K., Beer, D. G., Pennathur, A., Raphael, B. J., Wu, H.-T., Odze, R., Kim, H. K., Bowen, J., Leraas, K. M., Lichtenberg, T. M., Weaver, S., McLellan, M., Wiznerowicz, M., Sakai, R., Getz, G., Sougnez, C., Lawrence, M. S., Cibulskis, K., Lichtenstein, L., Fisher, S., Gabriel, S. B., Lander, E. S., Ding, L., Niu, B., Ally, A., Balasundaram, M., Birol, I., Bowlby, R., Brooks, D., Butterfield, Y. S. N., Carlsen, R., Chu, A., Chu, J., Chuah, E., Chun, H.-J. E., Clarke, A., Dhalla, N., Guin, R., Holt, R. A., Jones, S. J. M., Kasaian, K., Lee, D., Li, H. A., Lim, E., Ma, Y., Marra, M. A., Mayo, M., Moore, R. A., Mungall, A. J., Mungall, K. L., Nip, K. M., Robertson, A. G., Schein, J. E., Sipahimalani, P., Tam, A., Thiessen, N., Beroukhim, R., Carter, S. L., Cherniack, A. D., Cho, J., Cibulskis, K., DiCara, D., Frazer, S., Fisher, S., Gabriel, S. B., Gehlenborg, N., Heiman, D. I., Jung, J., Kim, J., Lander, E. S., Lawrence, M. S., Lichtenstein, L., Lin, P., Meyerson, M., Ojesina, A. I., Pedamallu, C. S., Saksena, G., Schumacher, S. E., Sougnez, C., Stojanov, P., Tabak, B., Taylor-Weiner, A., Voet, D., Rosenberg, M., Zack, T. I., Zhang, H., Zou, L., Protopopov, A., Santoso, N., Parfenov, M., Lee, S., Zhang, J., Mahadeshwar, H. S., Tang, J., Ren, X., Seth, S., Yang, L., Xu, A. W., Song, X., Pantazi, A., Xi, R., Bristow, C. A., Hadjipanayis, A., Seidman, J., Chin, L., Park, P. J., Kucherlapati, R., Akbani, R., Ling, S., Liu, W., Rao, A., Weinstein, J. N., Kim, S.-B., Lee, J.-S., Lu, Y., Mills, G., Laird, P. W., Hinoue, T., Weisenberger, D. J., Bootwalla, M. S., Lai, P. H., Shen, H., Triche, T., Van Den Berg, D. J., Baylin, S. B., Herman, J. G., Getz, G., Chin, L., Liu, Y., Murray, B. A., Noble, M. S., Askoy, B. A., Ciriello, G., Dresdner, G., Gao, J., Gross, B., Jacobsen, A., Lee, W., Ramirez, R., Sander, C., Schultz, N., Senbabaoglu, Y., Sinha, R., Sumer, S. O., Sun, Y., Weinhold, N., Thorsson, V., Bernard, B., Iype, L., Kramer, R. W., Kreisberg, R., Miller, M., Reynolds, S. M.,

Rovira, H., Tasman, N., Shmulevich, I., Ng, S. C. S., Haussler, D., Stuart, J. M., Akbani, R., Ling, S., Liu, W., Rao, A., Weinstein, J. N., Verhaak, R. G. W., Mills, G. B., Leiserson, M. D. M., Raphael, B. J., Wu, H.-T., Taylor, B. S., Black, A. D., Bowen, J., Carney, J. A., Gastier-Foster, J. M., Helsel, C., Leraas, K. M., Lichtenberg, T. M., McAllister, C., Ramirez, N. C., Tabler, T. R., Wise, L., Zmuda, E., Penny, R., Crain, D., Gardner, J., Lau, K., Curely, E., Mallery, D., Morris, S., Paulauskis, J., Shelton, T., Shelton, C., Sherman, M., Benz, C., Lee, J.-H., Fedosenko, K., Manikhas, G., Potapova, O., Voronina, O., Belyaev, S., Dolzhansky, O., Rathmell, W. K., Brzezinski, J., Ibbs, M., Korski, K., Kycler, W., ŁaŸniak, R., Leporowska, E., Mackiewicz, A., Murawa, D., Murawa, P., Spychała, A., Suchorska, W. M., Tatka, H., Teresiak, M., Wiznerowicz, M., Abdel-Misih, R., Bennett, J., Brown, J., Iacocca, M., Rabeno, B., Kwon, S.-Y., Penny, R., Gardner, J., Kemkes, A., Mallery, D., Morris, S., Shelton, T., Shelton, C., Curley, E., Alexopoulou, I., Engel, J., Bartlett, J., Albert, M., Park, D.-Y., Dhir, R., Luketich, J., Landreneau, R., Janjigian, Y. Y., Kelsen, D. P., Cho, E., Ladanyi, M., Tang, L., McCall, S. J., Park, Y. S., Cheong, J.-H., Ajani, J., Camargo, M. C., Alonso, S., Ayala, B., Jensen, M. A., Pihl, T., Raman, R., Walton, J., Wan, Y., Demchok, J. A., Eley, G., Mills Shaw, K. R., Sheth, M., Tarnuzzer, R., Wang, Z., Yang, L., Zenklusen, J. C., Davidsen, T., Hutter, C. M., Sofia, H. J., Burton, R., Chudamani, S., Liu, J., "Comprehensive molecular characterization of gastric adenocarcinoma", Nature, 513, 202–209 (2014).

[55] Agrawal, N., Akbani, R., Aksoy, B. A., Ally, A., Arachchi, H., Asa, S. L., Auman, J. T., Balasundaram, M., Balu, S., Baylin, S. B., Behera, M., Bernard, B., Beroukhim, R., Bishop, J. A., Black, A. D., Bodenheimer, T., Boice, L., Bootwalla, M. S., Bowen, J., Bowlby, R., Bristow, C. A., Brookens, R., Brooks, D., Bryant, R., Buda, E., Butterfield, Y. S. N., Carling, T., Carlsen, R., Carter, S. L., Carty, S. E., Chan, T. A., Chen, A. Y., Cherniack, A. D., Cheung, D., Chin, L., Cho, J., Chu, A., Chuah, E., Cibulskis, K., Ciriello, G., Clarke, A., Clayman, G. L., Cope, L., Copland, J., Covington, K., Danilova, L., Davidsen, T., Demchok, J. A., DiCara, D., Dhalla, N., Dhir, R., Dookran, S. S., Dresdner, G., Eldridge, J., Eley, G., El-Naggar, A. K., Eng, S., Fagin, J. A., Fennell, T., Ferris, R. L., Fisher, S., Frazer, S., Frick, J., Gabriel, S. B., Ganly, I., Gao, J., Garraway, L. A., Gastier-Foster, J. M., Getz, G., Gehlenborg, N., Ghossein, R., Gibbs, R. A., Giordano, T. J., Gomez-Hernandez, K., Grimsby, J., Gross, B., Guin, R., Hadjipanayis, A., Harper, H. A., Hayes, D. N., Heiman, D. I., Herman, J. G., Hoadley, K. A., Hofree, M., Holt, R. A., Hoyle, A. P., Huang, F. W., Huang, M., Hutter, C. M., Ideker, T., Iype, L., Jacobsen, A., Jefferys, S. R., Jones, C. D., Jones, S. J. M., Kasaian, K., Kebebew, E., Khuri, F. R., Kim, J., Kramer, R., Kreisberg, R., Kucherlapati, R., Kwiatkowski, D. J., Ladanyi, M., Lai, P. H., Laird, P. W., Lander, E., Lawrence, M. S., Lee, D., Lee, E., Lee, S., Lee, W., Leraas, K. M., Lichtenberg, T. M., Lichtenstein, L., Lin, P., Ling, S., Liu, J., Liu, W., Liu, Y., LiVolsi, V. A., Lu, Y., Ma, Y., Mahadeshwar, H. S., Marra, M. A., Mayo, M., McFadden, D. G., Meng, S., Meyerson, M., Mieczkowski, P. A., Miller, M., Mills, G., Moore, R. A., Mose, L. E., Mungall, A. J., Murray, B. A., Nikiforov, Y. E., Noble, M. S., Ojesina, A. I., Owonikoko, T. K., Ozenberger, B. A., Pantazi, A., Parfenov, M., Park, P. J., Parker, J. S., Paull, E. O., Pedamallu, C. S., Perou, C. M., Prins, J. F., Protopopov, A., Ramalingam, S. S., Ramirez, N. C., Ramirez, R., Raphael, B. J., Rathmell, W. K., Ren, X., Reynolds, S. M., Rheinbay, E., Ringel, M. D., Rivera, M., Roach, J., Robertson, A. G., Rosenberg, M. W., Rosenthall, M., Sadeghi, S., Saksena, G., Sander, C., Santoso, N., Schein, J. E., Schultz, N.,

Schumacher, S. E., Seethala, R. R., Seidman, J., Senbabaoglu, Y., Seth, S., Sharpe, S., Mills Shaw, K. R., Shen, J. P., Shen, R., Sherman, S., Sheth, M., Shi, Y., Shmulevich, I., Sica, G. L., Simons, J. V., Sipahimalani, P., Smallridge, R. C., Sofia, H. J., Soloway, M. G., Song, X., Sougnez, C., Stewart, C., Stojanov, P., Stuart, J. M., Tabak, B., Tam, A., Tan, D., Tang, J., Tarnuzzer, R., Taylor, B. S., Thiessen, N., Thorne, L., Thorsson, V., Tuttle, R. M., Umbricht, C. B., Van Den Berg, D. J., Vandin, F., Veluvolu, U., Verhaak, R. G. W., Vinco, M., Voet, D., Walter, V., Wang, Z., Waring, S., Weinberger, P. M., Weinstein, J. N., Weisenberger, D. J., Wheeler, D., Wilkerson, M. D., Wilson, J., Williams, M., Winer, D. A., Wise, L., Wu, J., Xi, L., Xu, A. W., Yang, L., Yang, L., Zack, T. I., Zeiger, M. A., Zeng, D., Zenklusen, J. C., Zhao, N., Zhang, H., Zhang, J., Zhang, J. (Julia), Zhang, W., Zmuda, E., Zou., L., "Integrated Genomic Characterization of Papillary Thyroid Carcinoma", Cell, 159, 676–690 (2014).

[56] Levine, D. A., "Integrated genomic characterization of endometrial carcinoma", Nature, 497, 67–73 (2013).

[57] Goy, G., Gezer, C., Gungor, V. C., "Credit Card Fraud Detection with Machine Learning Methods", 2019 4th International Conference on Computer Science and Engineering (UBMK), Samsun, Turkey, (2019).

[58] Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B., "KNIME: The Konstanz Information Miner", Berlin, Heidelberg, (2008).

[59] "R: The R Project for Statistical Computing", https://www.r-project.org/ (17.04.2021).

[60] Kolde, R., Laur, S., Adler, P., Vilo, J., "Robust rank aggregation for gene list integration and meta-analysis", Bioinformatics, 28, 573–580 (2012).

[61] Meyer, P., "Information-Theoretic Variable Selection and Network Inference from Microarray Data", Université Libre de Bruxelles (ULB), Brussels, Belgium, (2008).

[62] Hand, D. J., "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems", Machine Learning, 45, 171–186 (2004).

[63] Goy, G., Yazici, M. U., Bakir-Gungor, B., "A New Method to Identify Affected Pathway Subnetworks and Clusters in Colon Cancer", 2019 4th International Conference on Computer Science and Engineering (UBMK), Samsun, Turkey, (2019).

[64] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T., "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks", Genome Research, 13, 2498–2504 (2003).

[65] Yang, Z., Ren, F., Liu, C., He, S., Sun, G., Gao, Q., Yao, L., Zhang, Y., Miao, R., Cao, Y., Zhao, Y., Zhong, Y., Zhao, H., "DbDEMC: a database of differentially expressed miRNAs in human cancers", BMC Genomics, 11, 5 (2010).

[66] Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G., Liu, Y., "MiR2Disease: a manually curated database for microRNA deregulation in human disease", Nucleic Acids Research, 37, D98–D104 (2009).

[67] Xie, B., Ding, Q., Han, H., Wu, D., "MiRCancer: a microRNA-cancer association database constructed by text mining on literature", Bioinformatics, 29, 638–644 (2013).

[68] Ruepp, A., Kowarsch, A., Schmidl, D., Buggenthin, F., Brauner, B., Dunger, I., Fobo, G., Frishman, G., Montrone, C., and Theis, F. J., "PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes", Genome Biology, 11, 6 (2010).

[69] Huang, Z., Shi, J., Gao, Y., Cui, C., Zhang, S., Li, J., Zhou, Y., Cui, Q., "HMDD v3.0: a database for experimentally supported human microRNA–disease associations", Nucleic Acids Research, 47, D1013–D1017 (2019).

[70] Seo, J.-S., Ju, Y. S., Lee, W.-C., Shin, J.-Y., Lee, J. K., Bleazard, T., Lee, J., Jung, Y. J., Kim, J.-O., Shin, J.-Y., Yu, S.-B., Kim, J., Lee, E.-R., Kang, C.-H., Park, I.-K., Rhee, H., Lee, S.-H., Kim, J.-I., Kang, J.-H., Kim, Y. T., "The transcriptional landscape and mutational profile of lung adenocarcinoma", Genome Research, 22, 2109–2119 (2012).

[71] Kozomara, A., Birgaoanu, M., Griffiths-Jones, S., "MiRBase: from microRNA sequences to function", Nucleic Acids Research, 47, D155–D162 (2019).

[72] Bandyopadhyay, S., Mitra, R., Maulik, U., Zhang, M. Q., "Development of the human cancer microRNA network", Silence, 1, 6 (2010).

[73] Feng, Y.-H. Tsao, C.-J., "Emerging role of microRNA-21 in cancer", Biomedical Reports, 5, 395–402 (2016).

[74] Faragalla, H., Youssef, Y. M., Scorilas, A., Khalil, B., White, N. M. A., Mejia-Guerrero, S., Khella, H., Jewett, M. A. S., Evans, A., Lichner, Z., Bjarnason, G., Sugar, L., Attalah, M. I., Yousef, G. M., "The Clinical Utility of miR-21 as a Diagnostic and Prognostic Marker for Renal Cell Carcinoma", The Journal Of Molecular Diagnostics, 14, 385–392 (2012).

[75] Gaudelot, K., Gibier, J.-B., Pottier, N., Hémon, B., Van Seuningen, I., Glowacki, F., Leroy, X., Cauffiez, C., Gnemmi, V., Aubert, S., Perrais, M., "Targeting miR-21 decreases expression of multi-drug resistant genes and promotes chemosensitivity of renal carcinoma", Tumor Biology, 39 (2017).

[76] Emami, S., Nekouian, R., Akbari, A., Faraji, A., Abbasi, V., and Agah, S., "Evaluation of circulating miR-21 and miR-222 as diagnostic biomarkers for gastric cancer", Journal Of Cancer Research And Therapeutic**s**, 15, 115-119 (2018).

Curriculum Vitae

2012 – 2017	B.Sc., Computer Engineering, Erciyes University, Kayseri, TURKEY
2018 – 2021	M.Sc., Electrical and Computer Engineering, Abdullah Gül University, Kayseri, TURKEY
2018 – 2021	Research Assistant., Electrical and Computer Engineering, Abdullah Gül University, Kayseri, TURKEY
2021 – Present	Scientific Programs Assistant Expert., The Scientific and Technological Research Council of Turkey, Ankara, TURKEY

SELECTED PUBLICATIONS AND PRESENTATIONS

J1) B. Kolukisa, H. Hacilar, G. Goy, B. Bakir-Gungor, A. Aral, V. C. Gungor, Diagnosis of Coronary Heart Disease via Classification Algorithms and a New Feature Selection Methodology – International Journal of Data Mining Science (May 2019)

J2) G. Goy, B. Kolukisa, B. Bakir-Gungor, I. Ugur, V. C. Gungor, Weighted Association Rules and Scoring Methodology for Cardiovascular Diseases- International Journal of Bioscience, Biochemistry and Bioinformatics (Oct. 2019)

J3) M. Yousef, B. Bakir-Gungor, A. Jabeer, G. Goy, R. Qureshi, L.C. Showe, Recursive Cluster Elimination based Rank Function (SVM-RCE-R) implemented in KNIME – F1000Research (Jan. 2021)

J4) G. Goy, M. Yousef, R. Mitra, C.M. Eischen, A. Jabeer, B. Bakir-Gungor, miRcorrNet: Machine learning-based integration of miRNA and mRNA Expression Profiles, combined with Feature Grouping and Ranking – PeerJ (May 2021)

J5) A Pathway and Network Oriented Approach to Enlighten Molecular Mechanisms of Type 2 Diabetes Using Multiple Association Studies – bioRxiv (Under Review)

J6) G. Goy, M. Yousef, B. Bakir-Gungor, miRModuleNet: Detecting miRNA-mRNA Regulatory Modules – Frontiers In Genetics (Under Review)

C1) B. Kolukisa, H. Hacilar, G. Goy, B. Bakir-Gungor, A. Aral, V. C. Gungor, Evaluation of Classification Algorithms, Linear Discriminant Analysis and a New Hybrid Feature Selection Methodology for the Diagnosis of Coronary Artery Disease – IEEE International Conference on Big Data (Big Data) (Dec. 2018)

C2) G. Goy, M.U. Yazici, B. Bakir-Gungor, A New Method to Identify Affected Pathway Subnetworks and Pathway Clusters in Colon Cancer – 4th International Conference on Computer Science and Engineering (Sep. 2019)

C3) G. Goy, C. Gezer, V. C. Gungor, Credit Card Fraud Detection with Machine Learning Methods – 4th International Conference on Computer Science and Engineering (Sep. 2019)

C4) G. Goy, B. Kolukisa, C. Bahcevan, V. C. Gungor, Ensemble Churn Prediction for Internet Service Provider with Machine Learning Techniques – 5th International Conference on Computer Science and Engineering (Sep. 2020)