Indian Statistical Institute

Doctoral Thesis

Developing Pattern Recognition and Interpretable Convolutional Neural Network based Frameworks for identifying drug resistant and pan cancer miRNAs from Expression Data

A thesis submitted to the Indian Statistical Institute
in partial fulfilment of the requirements for
the degree of
Doctor of Philosophy (in Computer Science)

Author:
Joginder Singh

Supervisor:
Prof. Shubhra Sankar Ray

Center for Soft Computing Research Unit

Indian Statistical Institute, Kolkata

Friday 16th January, 2026

Dedicated to My beloved Parents.

Acknowledgements

This thesis is the end of my journey in obtaining Ph.D. This thesis is completed with the support and encouragement of numerous people, including my colleagues, well-wishers, friends, and relatives, and all those people who made this thesis possible. Few words can't express the immeasurable value of support I received. Thanksgiving is just a formality, and I don't know how to express my gratitude to all my beloved well-wishers.

First and foremost, I would like to thank my supervisor, Professor Shubhra Sankar Ray, without whose guidance, encouragement, and support, this thesis could not have been completed. He helped me to understand the problems related to microarray expression and guided me to use computational skills to solve those problems. Thanks are due to him for his kind permission to include the joint research work in this thesis. I would also like to express my gratitude to Prof. Sankar Kumar Pal, Dr. Kuntal Ghosh, and Prof. Ashish Ghosh for their advice and suggestions.

I am thankful to my friends Ms. Srijanie Banerjee, Dr. Jayanta Kumar Pal, Ms. Sukriti Roy, Ms. Navpreet Kaur, Ms. Amrita Kundu, Mr. Keerthi S. Chandran, Dr. Sankar Mondal, Dr. Susanta Samanta and Dr. Bibhuti Das. I am also grateful to my labmates Dr. Anjan Chowdhury, Ms. Sandipa Roy, Ms. Barnini Bhattacharya, Dr. Debashree Dutta, and Mr. Shibsankar Roy. I am also thankful to Mr. Sujit Basak, Mrs. Pamli Sengupta, Mrs. Tapashi Srimani, and the Dean's office staff for helping me in many ways with all official work-related problems. I would also like to express my gratitude to the RFAC and PhD/DSc. committee members for helping me.

In every person's life, there is always a group of people who help them grow as a better person, build their character, and maintain consistency and perseverance in work. I feel blessed to have such friends in my life and want to thank Sohini Boral, Manish Kumar, Sandeep Kajal, Deepak Gothwal, Ankit Dhaka, Bhuvan Deep, Srijanie Banerjee, and Indu Bala for helping me in the last few years.

Finally, I would like to express my sincere gratitude towards the most important people in my life, my family members. I would not have been able to complete my thesis work without the continuous support of my mother Smt. Birhma Devi and father Shri. Ganga Bishan (Late). I would also like to thank my brothers, Satish, Karambir, and Satbinder and their families. I am also grateful to my sister and her family for their support.

Joginder Singh

Friday 16th January, 2026

List of Related Publications

J1. A. Kundu, J. Singh, J. K. Pal and S. S. Ray "Predicting Drug-Resistant miRNAs in Cancer." ***Network Modeling Analysis in Health Informatics and Bioinformatics***, vol. 12, no. 6, pp. 1-17, 2023, DOI:`https://doi.org/10.1007/s13721-022-00398-8`. (Publisher: Springer.)

J2. J. Singh and S. S. Ray "Integrating fuzzy rough set-based entropies for identifying drug-resistant miRNAs in cancer." ***Journal of Computational Science*** vol. 91, pp. 102673, 2025, DOI:`https://doi.org/10.1016/j.jocs.2025.102673`. (Publisher: Elsevier.)

J3. J. Singh, S. S. Ray and S. Roy "Identifying Pan-cancer and Cancer Subtype miRNAs using Interpretable Convolutional Neural Network." ***Journal of Computational Science*** vol. 85, pp. 102510, 2025, DOI:`https://doi.org/10.1016/j.jocs.2024.102510`. (Publisher: Elsevier.)

J4. J. Singh and S. S. Ray "Multi-Objective Framework for Optimizing CNN and Set-theoretic Explainable AI based Attribution Score for Selecting miRNAs in Pan-cancer Data." **Status: Major Revision in Applied Intelligence (Manuscript ID: APIN-D-25-07047R2). Publisher: Springer**

Abstract

Micro Ribonucleic Acids (miRNAs) are short length (∼24) non-coding RNAs and are considered as key biomarkers in cancer diagnosis and treatment. They play a vital role in classifying cancer patients from normal ones and drug resistant patients from control ones. The control patients are those who have not received any drug for cancer treatment. The objective is to identify a subset of miRNAs those help in the classification of the patients using expression data. The thesis is comprised of four contributory chapters in addition to an introduction and conclusion. In the first two contributory chapters, computational methods for ranking and selecting miRNAs associated with drug resistance in cancer are introduced. In the fourth and fifth chapters, deep learning based methods are presented for identifying miRNAs for various cancer classes in pan-cancer data. The contributory chapters are as follows:

- Selecting drug-resistant miRNAs in cancer using Euclidean distance with fold change based score.
- Integrating fuzzy rough set-based entropies for identifying drug resistant miRNAs and classifying cancer patients.
- Interpretable convolutional neural network for selecting miRNAs from multiple cancer classes and cancer subtypes through pan-cancer analysis.
- Set-theoretic explainable AI-based attribution score for identifying miRNAs in pan-cancer data.

In Chapter 1, an introduction to the related problems, literature review, motivation, and the organization of the thesis are provided. In Chapter 2, two methods to predict the miRNAs associated with drug resistance in cancer are presented. While, in the first method, a score is developed using the Euclidean distance with weighted fold change (EDWFC), in the second method, a histogram-based clustering and Euclidean distance with fold change-based ranking (HCEDFCR) is introduced. The EDWFC provides a ranked list of miRNAs for classifying control and drug-resistant patients and the HCEDFCR returns a group of miRNAs associated with drug resistance. The methods are trained with the help of existing biological knowledge. In Chapter 3, two new z score based fuzzy rough relevance and redundancy entropies are developed, and then a weighted framework is introduced to integrate the entropies for ranking and selecting miRNAs. The selected miRNAs are used for classifying the control and drug-resistant patients. In Chapter 4, an interpretable one dimensional convolutional neural network

model (ICNNM) is developed and it is optimized in terms of hyperparameters for identifying classes of patients among multiple cancer classes in pan-cancer data. An attribution scores is also introduced using SHapley Additive exPlanations (SHAP) values for interpreting the miRNAs and selecting important miRNAs for each cancer class. In Chapter 5, a multi-objective framework for optimizing hyperparameters of a 1D CNN, called MOHCNN, and a set-theoretic explainable AI-based attribution scores (STEAAS) for miRNA selection are developed. The objectives for optimization are training error, validation error, and the number of training parameters. A set-theoretic explainable AI-based attribution score is developed for identifying miRNAs in various cancers. The score of a miRNA is represented by an ordered pair, where the first part represents the class score of the miRNA, and the second part denotes the reliability score of that miRNA for belonging to the class. The miRNAs with high class scores and reliability scores in a class are selected.

All the developed methods are compared with related miRNA and gene selection techniques and popular classifiers. Data from public repositories such as Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) data) are used. The biological significance of the miRNAs, selected by the developed methods, is established using publicly available web based bioinformatics tools and existing literature.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

Cancer is a disease with a high mortality rate and it is a burden on public health worldwide [1]. Currently, approximately 20 million of patients are estimated to be suffering from cancer in the world [2]. The maximum number of patients is related to lung cancer. Cancer occurs when a cell starts abnormal cell division and begins to grow rapidly. This can happen almost anywhere in the body, and it can take years for an abnormal cell to grow into a tumor. When abnormal cells invade neighboring tissues or spread to other parts of the body through the blood and lymph systems, then the process is called metastasis. There are many different types of cancer such as sarcoma, leukemia, lymphoma, and multiple myeloma. Cancer types are divided into more than 100 classes and also subclasses, based on the location of the tumor, according to various literature [3]. For example, prostate cancer alone is divided into seven subclasses depending on the micro ribonucleic acid (miRNA) or gene regulations, fusions, and mutations [4]. In general, cancer treatment is difficult as most people are diagnosed when the cancer is spread to the whole tissue or other nearby tissues [5]. Early diagnosis and precise drug treatment can help in lowering the mortality rate of cancer patients. In this regard, miRNAs are considered as one of the key biomarkers in cancer identification and drug treatment for patients [6–8]. MiRNAs can be collected from tissue, blood, and saliva of patients, and the expression of miRNAs can be studied to detect cancer and its type. The expression data can be handled computationally to identify its patterns in various cancers.

For an individual cancer type, cancer analysis can be performed by comparing the miRNA expressions of patients with normal persons. For handling multiple classes of cancers and identifying class specific miRNAs and patients, one can analyze expressions of various cancer types together which is called pan-cancer analysis. Further, during treatment, a person may develop resistance to certain drugs which is referred as drug resistance. This can be detected by observing the change in the expression of certain miRNAs from the control group of patients to drug resistant patients. While, the control

group consists of patients who have not received chemotherapy, the resistant group consists of patients who have received drug treatment and developed drug resistance.

The focus of this thesis is on i) the identification of drug resistant miRNAs using miRNA expression data of cancer patients who received drug treatment and ii) the detection of important miRNAs for various cancer types using pan-cancer expression data. Four contributory chapters are dedicated to achieve these objectives. As a result, this thesis is comprised of six chapters, including four contributory chapters (Chapters 2, 3, 4, and 5). In the first contributory chapter, the objective is to identify drug resistant miRNAs in various cancer types. To achieve this objective, two computational methods, called 'Euclidean distance and weighted fold change-based ranking' (EDWFC) and 'Histogram-based clustering and the Euclidean distance with fold change-based ranking' (HCEDFCR), are presented to identify drug resistant miRNAs in cancer. The novelty of the investigation lies in utilizing the biological knowledge of known drug resistant miRNAs for ranking the miRNAs and selecting the relevant ones. The purpose of Chapter 3 is to identify a set of miRNAs whose expression values can help in classifying the control and drug resistant patients efficiently. To achieve this, a weighted framework for integrating two new z score based relevance and redundancy entropies (WFIFRRRE) for miRNA (feature) selection and for classifying control and drug resistant patients is presented. The novelty of this work lies in formulating a new z score based fuzzy membership function to compute the membership of each expression value to calculate the relevance and redundancy entropy measures, and then developing a framework for integrating the relevance and redundancy entropies for miRNA ranking and selection. Chapters 4 and 5 are dedicated to find a set of important miRNAs for various cancer types using pan-cancer expression data. In Chapter 4, a one-dimensional (1D) interpretable convolutional neural network model (ICNNM) is developed for identifying relevant miRNAs in various cancer types and subtypes. The novelty of the method lies in using Bayesian optimization with multivariate tree parzen estimator (BoMTPE) for optimizing hyperparameters of a one-dimensional convolutional neural network model and integrating Shapley Additive exPlanations (SHAP) with the optimized CNN to compute attribution scores of miRNAs in different cancer types. The objective of Chapter 5 is to select miRNAs with reliability from various cancer classes in Pan-cancer Data by developing a set-theoretic explainable artificial intelligence-based attribution score for miRNAs. In the process a multi-objective framework for optimizing CNN is developed. The novelty of the framework lies in integrating two explainable AI-based scores, BayLIME and SHAP, and combining them with optimized CNN in multi-objective framework for identifying relevant miRNAs with reliability.

1.1 Basics of MiRNA

A miRNA is a short-length (∼22nt) non-coding RNA extracted from ∼60-110nt RNA precursor [9] and regulates gene expression by binding to a particular messenger RNA (mRNA) and promoting their degradation and/or translational inhibition [10]. The number of miRNAs in most organisms is relatively lower than the number of genes (e.g., a human genome is considered to encode approximately 1000 miRNAs, whereas the estimated number of genes is approximately 30000). A single miRNA can regulate more than 100 genes, which can significantly affect the gene expression networks [10]. MiRNA expression profiles are considered to be rich in biological information, since changes in the expression of hundreds of genes may be reflected by identifying the change in the expression of one or a few miRNAs that may have regulated them [11].

Researchers found that changes in miRNA expression profiles play an important role in the development of cancer and drug resistance during treatment [12] that leads to disease progression, even after the application of drugs, and decreases life expectancy. Dysregulation of miRNAs controls the malignant cells and can act as a tumor suppressor or promoter. There are multiple reasons for miRNA dysregulation, such as mutation, deletions, and epigenetic changes [13].

Caline et al. [14] first identified the regulation of miRNAs, miR-15, and miR-16, in chronic lymphocytic leukemia (CLL) in 2002, and in the last two decades, miRNAs are established as one of the key components in cancer diagnosis and treatment. Researchers from various fields like medical science, biology, bioinformatics, chemistry, statistics, and computer science contributed on identifying the role of miRNAs in different cancers. Mitchell et al. [15] established that miRNAs identified in plasma or serum can be utilized for the detection of cancer. Furthermore, expression levels of miRNAs in serum can differentiate the healthy and cancer samples [16]. Subtype identification of cancer is also possible using miRNA expressions [8].

1.1.1 MiRNA Generation

MiRNA expressions are generated in biological laboratories and the process can be broken into two steps. The first step is the generation of short-length miRNA strands from the cell and then processing those strands in the laboratory to assess the quantity of expressed miRNAs.

MiRNA Biogenesis: As discussed earlier, miRNAs are short-length (∼22nt) RNAs and an example of miRNA biogenesis is considered from [10] and provided in Fig. 1.1 [1]. The four major steps to generate them are as follows:

[1]The figure is taken from [10] which is available publicly.

1. Generation of primary miRNA: MiRNAs are generated by transcribing miRNA-coding sequences into long primary transcripts (pri-mRNAs) [17].

2. Generation of Precursor miRNA: Pri-mRNA is cleaved by drosha and pasha enzymes in the nucleus and it results in a hairpin-looped precursor nucleotide (pre-miRNA) of approximately 75 nucleotides in length.

3. Transportation of Precursor miRNA into Cytoplasm: The pre-miRNA is then transported out of the nucleus into the cytoplasm by Exportin 5 [18].

4. Generation of mature strand miRNA: RNAse III enzyme, dicer, cleaves the pre-miRNA and generates approximately a ∼24 nucleotide long duplex structure with a mature and complementary passenger strand. The simplex mature miRNA strand is generated from the mature strand.

In Fig. 1.1, the generation of mature strand miRNA from primary miRNA (pri-mir) is shown. Pri-mir is transformed into pre-miRNA (pre-mir) by drosha enzyme in the nucleus. Pre-miRNA is then sent to the cytoplasm as a double-stranded miRNA, where the strand starting from 5 prime (5p) carbon is mentioned as miR* and the strand starting from 3p carbon is mentioned as miR. The double stranded miRNA is cleaved by the dicer enzyme into two strands where the 5p strand is degraded and the 3p strand is loaded into the RNA-induced silencing complex (RISC) [10].

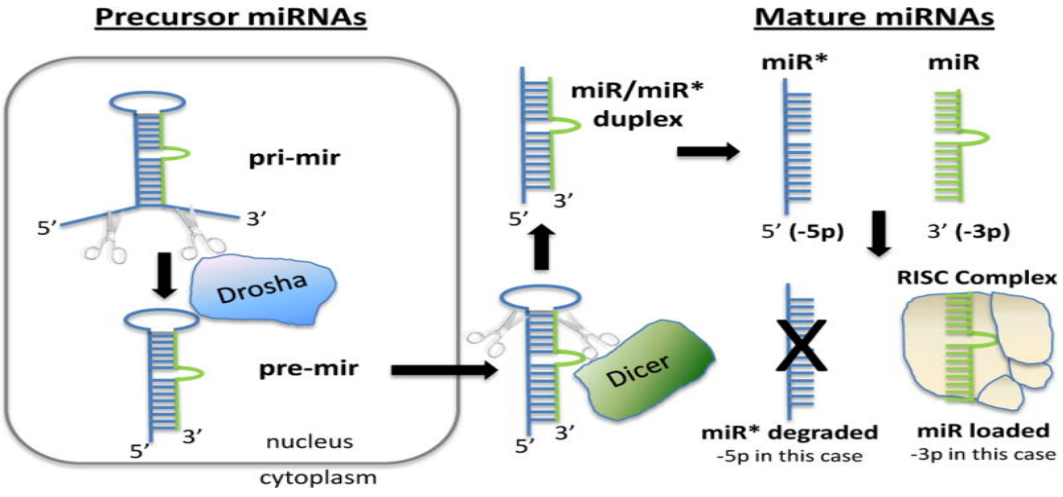


Figure 1.1: MiRNA biogenesis showing the preparation of mature miRNA from precursor miRNA.

In Table 1.1, the mature sequences of hsa-miR-155, hsa-miR-let-7e-5p, and hsa-miR-21-3p from 5p carbon to 3p carbon are shown, where these are extracted from the precursor miRNAs (pre-miRNAs) [19]. A miRNA is represented mainly using three parts. The first part denotes the species, the second represents that it is a miRNA, and the third part describes a unique identity number. The unique identity number of a miRNA denotes the order of its discovery and registration to the miRNA database, such as miRBase [19]. For example, in hsa-miR-155, hsa represents that it is a human

miRNA, miR implies that it is a micro RNA, and 155 implies that it is the 155th miRNA to be officially registered.

Table 1.1: Example of miRNA sequence

miRNA	**Sequence (5p to 3p)**
hsa-miR-155	CUCCUACAUAUUAGCAUUAACA
hsa-miR-let-7e-5p	UGAGGUAGUAGGUUGUAUAGUU
hsa-miR-21-3p	CAACACCAGUCGAUGGGCUGU

1.1.2 Expression Generation

After mature strand creation, the miRNAs are colored with fluorescent dye. The color intensity value of a miRNA can be scanned and saved as its expression [20]. This can be done in three major processes as follows:

1. Expression profiling by cloning and sequencing: It involves separating mature miRNA, adaptor binding, reverse transcription, and polymerase chain reaction (PCR) amplification. [21].

2. Microarray analysis: Microarray analysis includes design of oligonucleotide probes, preparation of labeled material from RNA samples (with or without amplification), and preparation of microarray [22].

3. Microbead analysis: Here, every miRNA is considered as a microbead. One effective approach to microbead expression analysis is xMAP, which allows for the simultaneous investigation of one hundred distinct miRNAs.

Table 1.2: Example of miRNA expression values in Drug resistant data.

miRNA	Drug resistant				Control		
	Patient1	Patient2	Patient3	Patient4	Sample1	Sample2	Sample3
miR-21	8.1	8.3	8.2	8.4	3.9	4.2	4.1
miR-34a	3.2	3.3	3.1	3.4	3.0	3.2	3.1
miR-155	6.5	6.7	6.4	6.6	6.3	6.2	6.4
miR-200c	5.9	6.1	5.8	6.0	5.7	5.9	5.8
miR-16	7.3	7.1	7.2	7.4	6.9	7.0	6.8
miR-145	5.0	4.9	5.1	5.2	4.8	4.9	5.0
miR-451	6.1	5.9	6.0	6.2	6.0	5.8	5.9
miR-125b	5.8	5.9	5.7	5.8	5.6	5.7	5.5

Table 1.3: Example of miRNA expression values in pan-cancer.

miRNA	Breast					Lung					Colon				
	S1	S2	S3	S4	S5	S1	S2	S3	S4	S5	S1	S2	S3	S4	S5
miR-21	8.2	8.5	8.3	8.1	8.4	9.1	8.9	9.0	9.2	9.3	9.4	9.2	9.1	9.3	9.5
miR-34a	3.1	3.4	3.3	3.2	3.1	2.8	3.0	3.1	3.2	2.9	2.9	3.2	3.0	3.1	3.3
miR-155	6.5	6.3	6.4	6.6	6.2	7.0	6.9	7.1	6.8	7.2	7.2	7.1	7.0	7.3	7.2
miR-200c	5.9	5.7	5.8	5.9	5.8	6.1	6.3	6.0	6.2	6.1	6.0	5.8	5.9	6.1	6.0
miR-16	7.2	7.1	7.3	7.2	7.0	6.8	7.0	6.9	7.1	6.9	6.9	7.0	6.8	7.1	6.9
miR-145	4.9	5.0	5.1	4.8	4.9	4.8	4.7	4.9	5.0	4.8	4.8	4.9	5.0	4.9	5.1
miR-451	6.0	5.8	5.9	5.7	5.8	6.2	6.1	6.3	6.0	6.1	6.3	6.2	6.1	6.4	6.3

The miRNA expression values can be represented as a data matrix where numerical values in the matrix cell represent the expression level of a miRNA in a patient suffering from a particular disease (in our case, cancer disease). The examples of miRNA expression values in drug resistant and pan-cancer datasets is provided in Table 1.2 and 1.3, respectively. For example, the expression values of miR-21 are 8.1, 8.3, 8.2, and 8.4 in Patient1, Patient2, Patient3, and Patient4, respectively, in the drug resistant class. The expression values for the same miRNA are 3.9, 4.2, and 4.1 in Sample1, Sample2, and Sample3, respectively, in the control class. In a similar way, the expression values of 7 miRNAs for 15 patients in breast, lung, and colon cancer classes is shown in Table 1.3.

The data matrix of miRNA expression values is utilized for analysis using statistical or machine learning methods to predict differentially expressed miRNAs, classify samples (e.g., diseased vs. healthy), or find target genes. The level of expression values of a miRNA represents the regulation of that miRNA for a certain disease. While a high expression value of a miRNA corresponds to strong expression in a disease, a low or near-zero expression value denotes no expression.

1.2 Preliminaries of Pattern Recognition-based Techniques

In this section, we discuss some preliminary concepts of pattern recognition that are helpful in miRNA selection, clustering, and patient classification.

1.2.1 Clustering

Clustering is an unsupervised learning process for grouping data points. In the miRNA expression domain, clustering helps to group the miRNAs or patients based on their expression profiles whose functions or disease conditions are expected to be similar. For example, miRNAs that are resistant to a specific drug can be grouped together based on their expression profiles. Some of the clustering methods are discussed below.

K-means Clustering [23–25]: K-means clustering partitions the data into K groups, where K is a user defined number of clusters. The concept of K-means relies on minimizing the variance among the data points in each cluster. The first step is the selection of K-random values as cluster centers. The second step is assigning each data point to the nearest cluster center. The third step is to compute the mean of all the data points assigned to each cluster as cluster center, update the previous cluster center with the new one, and compute the variance among the data points for each cluster. The second and third steps are repeated until the variance among the data points in each cluster remains unchanged.

Spectral Clustering [26–28]: Spectral clustering is a graph-based data partitioning algorithm in which each data point is represented as a node. Edges between all possible

pairs of nodes are calculated as an affinity matrix and it signifies the relationships, may be distances, among the data points. The affinity matrix is converted into a Laplacian matrix for computing the eigen values and eigen vectors such that communities, i.e., closely connected graphs, within the graph can be identified. Finally, clusters are formed based on the similarity among communities.

Self-Organization Maps (SOMs) [29–31]: SOMs are competitive learning-based neural networks that cluster the data points while maintaining the topological relationship among them. These are also used for data reduction and visualization into two-dimensional space such as hexagonal or rectangular grids. Two important parameters that control the learning of SOM during training are learning rate and neighborhood radius, which decrease over the iterations (a user defined number) to group the data points that are more similar to each other. The training of SOM involves four major steps: initialization, competition, cooperation, and adaptation. The first step is the initialization of k weight vectors (nodes) where K is the number of clusters, weights of vectors are initialized randomly, and the size of each weight vector is the same as the size of each data point. The second step is the computation of the distance between a data point and each weight vector. The weights of the node are updated, which has a minimum distance from the data point, and that node is called the best matching unit (BMU). This is repeated for all data points. The third step is cooperation where the weights of BMU and its neighboring nodes are updated to bring them more closer to input data points. The fourth step is adaptation, which helps decrease the number of neighborhood data points that possess lesser similarity with the points in the cluster. The steps from 2-4 are repeated until the weights of vectors stop updating.

1.2.2 Feature Selection

Most of the datasets contain irrelevant features that may result in poor performance of machine learning models used for analysis. To improve the performance of a model, the removal of irrelevant features is crucial and it can be handled using feature selection techniques. These methods identify the important subset of features that help in classifying or grouping samples. Feature selection can be categorized as follows:

Filter Methods: These methods select the features based on their relations such as correlation, mutual information, class difference, etc. However, different relations can be used together to determine a subset of features. For example, if someone wants to have multiple groups of features with high class differences and maximum correlation within the group, then a distance metric can be used with a correlation measure. These features are selected before using any machine learning model which makes filter methods computationally efficient. Some methods based on the filtering approach are combination of Fisher score, relief F, and mutual information [32], minimum redundancy and maximum

relevance [33], fold change-based ranking [34], fuzzy rough entropy measure [35], fuzzy mutual information based measure [36], and set theoretic entropy measure [37].

Wrapper Methods: Wrapper methods determine an optimal set or group of features with the help of machine learning models such as classifiers or clustering methods. For example, if a classifier is used then the classification accuracy is stored for each subset of features, and the subset of features responsible for maximum classification accuracy is considered as the optimal set of features. For clustering methods, those groups of features are considered important which provide better clustering index values or low variance among data points within the cluster. Some techniques based on the wrapper method are recursive feature elimination [38], sequential forward selection [39], sequential backward selection [39], sequential backward floating selection [39], and Covariance Matrix Adaptation Evolution [40],

Hybrid Methods: These methods consider both filter and wrapper techniques in parallel and combine their advantages. In general, while the filter method selects features based on their properties, the wrapper method updates the feature set based on the performance of the classifier. Some of the hybrid approaches are uncertainty measures using neighborhood entropy [41], combination of Adaptive Genetic Algorithm and Mutual Information Maximization [42], optimizing fuzzy mutual information using particle swarm optimization [43], correlation based gene selection using genetic algorithm [44], and parallelized correlation based gene subset selection [45].

1.2.3 Classification

Classification is a supervised learning process for assigning labels to samples or data points. This is achieved using classifiers that require the labeled samples for training and then classifying them accurately in the test process. Using this training, the classifier then predicts the labels of the unknown samples. Therefore, the training of the classifier should be done in a way that the classifier identifies the important patterns and neglects the noise in the data during training. Further, the performance of any classifier depends on the features selected by any feature selection technique. Some of the popular classifiers are as follows:

Support Vector Machines [46–48]: Support vector machine uses the concept of hyperplanes and uses kernel function/s to partition the data points into different classes. The kernel functions are helpful in determining the relationships such as linear, nonlinear, shape based, topological, etc., among data points. Hence, the classification performance of SVM varies for different kernel functions. The classification process of SVM involves finding an optimal hyperplane, maximizing the margins, and adjusting support vectors. The objective is to maximize the margin between support vectors

and the hyperplane, where support vectors are the nearest data points to the hyperplane. For example, consider a dataset with two classes and n patients represented as $(X_1, y_1), (X_2, y_2), (X_3, y_3), \ldots, (X_n, y_n)$, where $X_i \in R^d$ is a sample containing d real-valued features and $y_i \in \{-1, 1\}$ denotes the class label of patients. The goal is to find a hyperplane $W^T X + b = 0$ that separates the classes with a maximum margin. This can be formulated as:

$$\arg\max_{w,b} \frac{1}{2}||W||^2 \tag{1.1}$$

subject to

$$y_i(W^T X_i + b) \geq 1 \ \forall i \tag{1.2}$$

Eq. 1.1 guarantees the maximum margin between the support vectors and the hyperplane. The term $\frac{1}{2}||W||^2$ is minimized to maximize the margin.

Naive Bayes [49,50]: It utilizes the concept of Bayes theorem for categorizing samples in their respective classes. It is suitable for both binary and multiclass classification. The mathematical expression of Bayes theorem is as follows:

$$P(C|X) = \frac{P(X|C) \times P(C)}{P(X)} \tag{1.3}$$

where $P(C|X)$, $P(X|C)$, $P(X)$, $P(C)$ represent the probability of the occurrence of, class C given sample X, sample X given class C, class C, and sample X, respectively. X is denoted by n features $X = [x_1, x_2, x_3, \ldots, x_n]$. In naive bayes, every feature is considered to be independent. The goal is to determine the posterior probability $P(C|X)$ by utilizing the relation of joint and conditional probabilities.

$$\begin{aligned} P(X|C) &= P(x_1, x_2, x_3, \ldots, x_n, C), \\ &= P(x_1|x_2, x_3, \ldots, x_n, C) \times P(x_2, x_3, \ldots x_n, C) \\ &= P(x_1|x_2, x_3, \ldots, x_n, C) \times P(x_2|x_3, \ldots x_n, C) \times P(x_3, x_4, \ldots x_n, C) \\ &= P(x_1|x_2, x_3, \ldots, x_n, C) \times \cdots \times P(x_{n-1}|x_n, C) \times P(x_n, C) \end{aligned} \tag{1.4}$$

With the assumptions of independent features, $P(x_i|x_{i+1}, \ldots, x_n, C)$ is reduced to $P(x_i, C)$. Now, Eq. 1.3 can be written as:

$$P(C|X) = \frac{\Pi_{i=1}^{n} P(x_i|C) \times P(C)}{P(X)} \tag{1.5}$$

Consider an example of two classes (C_1, C_2). Based on Eq. 1.5, the sample X can be assigned to class C_1 if $P(C_1|X) \geq P(C_2|X)$ else to class C_2 .

Random Forest [50,51]: Random forest is a tree-based classifier that uses an ensemble of various decision trees to increase the classification accuracy. The predictions of different trees are aggregated by using an average of the values or voting algorithm.

The limitations of individual trees such as high variance and instability of features are tackled by exploiting the concepts of feature randomness and bagging. The main steps of random forest are as follows:

Ensemble of Decision Trees: Random forest divides the datasets into many regions randomly and various regions are then used to train different decision trees independently. Predictions from all the trees are combined using selection procedures such as voting.

Bootstrap Aggregating: Random subsets of features are used with replacements to train the trees in the forest. This process helps in some features getting selected many times and some being left out, and that helps in preventing overfitting by decreasing the variance of the model.

Feature Randomness: Random forest only considers a subset of random features while splitting the nodes of trees instead of all the features like a decision tree. This increases the robustness and generalization of the model by decreasing the correlation among trees.

Out-of-Bag (OOB) Error: Each tree is trained with a bootstrapped sample, thus a part of the data is only selected. The excluded samples, referred to as out-of-bag samples, are utilized to assess the model's performance without requiring a distinct validation set.

1.3 Performance Evaluation

The performance evaluation techniques are useful in determining the efficiency of selected features while classifying samples. The learning process of any model is divided into two major parts training and testing. The training part utilizes the available information about samples to train the model. The testing part involves the decision provided by the model for unseen samples. The developed methodologies in this thesis are evaluated in terms of various measures such as sensitivity, specificity, F-score etc., using cross-validation (CV) techniques for various classifiers. In the next two sections, the performance measures and cross-validation techniques are discussed.

1.3.1 Performance Measures

The performance measures help in evaluating the performance of a classifier by observing the correct and wrong predictions. A confusion matrix is constructed that shows the predictions using true positive, false positive, false negative, and true negative performance measures. The rows in the confusion matrix represent the predicted labels of patients by the classifier, and the columns denote the actual labels of the same (See Fig. 1.2). The positives and negatives in the matrix are assigned based on the class labels.

For example, let us consider the $C1$ class as positive and $C2$ as negative. Then the variables can be defined as:

True Positive (TP): When the predicted label $C1$ of patients aligns with the actual class label $C1$.

False Positive (FP): The patients from class $C2$ are predicted as patients from class $C1$.

False Negative (FN): The patients from class $C1$ are predicted as patients from class $C2$.

True Negative (TN): When the patients from class $C2$ are predicted as patients from class $C2$.

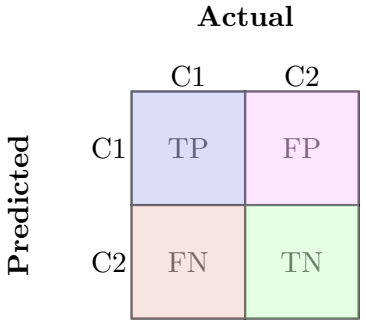


Figure 1.2: Confusion Matrix for binary class.

The above variables are utilized to compute the performance measures such as sensitivity, specificity, F1-score (F score), accuracy, and Matthews correlation coefficient (MCC) for evaluating the developed methods. The measures are defined as follows:

Sensitivity/Recall: It measures the ratio between the number of correctly predicted positive patients and the number of actual positive patients. It is defined as:

$$Sensitivity = \frac{TP}{TP + FN} \tag{1.6}$$

Specificity: It measures the ratio between the number of correctly predicted negative patients and the number of actual negative patients. It is defined as:

$$Specificity = \frac{TN}{TN + FP} \tag{1.7}$$

Precision: It measures the ratio between the number of truly predicted positive patients and the number of all predicted positive patients. It is defined as:

$$Specificity = \frac{TP}{TP + FP} \tag{1.8}$$

F1-score: It measures the accuracy of a classifier by computing the harmonic mean of precision and sensitivity (recall). It is defined as:

$$F1 - score = \frac{2 \times TP}{2 \times TP + FN + FP} \tag{1.9}$$

Accuracy: It measures the capability of a classifier in assigning samples to their correctly labeled classes.

$$Accuracy = \frac{Correctly\,classified\,samples}{total\,number\,of\,samples} * 100 \tag{1.10}$$

MCC: It measures the correlation between predictions and actual labels. It is defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{1.11}$$

The sensitivity, specificity, and F1-score values range from 0 to 1. Accuracy can be denoted in the range of 0 to 1 (in ratio) and 0 to 100 (in percentage). The MCC ranges from -1 to +1 which indicates whether the prediction capability is better or worse than random prediction. A positive value of MCC denotes that the prediction capability is better than random prediction.

The pan-cancer data analysis is a multiclass problem. The positives and negatives cannot be assigned in multiclass classification like binary classes. Here, the positives and negatives are computed for each class, and to find out the overall performance of the model, their average value is determined. This is better understood by an example provided in Fig. 1.3. In the figure, the cells with colors blue, green, purple, and brown represent true positive (TP), true negative (TN), false positive (FP), and false negative (FN), respectively, for a class from multiple classes.

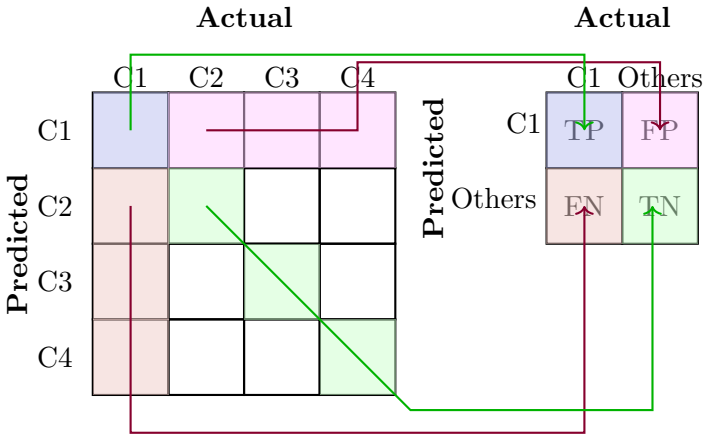


Figure 1.3: Computation of positives and negatives for a class from confusion matrix for multiclass problem.

The above measures are defined for one class. Similarly, these can be computed for the remaining classes such as C2, C3, and C4. The F1-score for class C1 can be computed as:

$$F1 - score_{C1} = \frac{2 \times TP_{C1}}{2 \times TP_{C1} + FN_{C1} + FP_{C1}} \tag{1.12}$$

Similarly, other measures can also be computed. The overall F-score for the model is defined as:

$$F1 - score = \frac{F1 - score_{C1} + F1 - score_{C2} + \cdots + F1 - score_{CN}}{Number\ of\ classes} \tag{1.13}$$

1.3.2 Cross Validation

Cross validation (CV) is a prediction evaluation process to enhance the robustness of a computational model while applying it to various types of datasets [52]. Sometimes a model cannot identify the patterns in the data and it results in random predictions or lower accuracy which is considered as underfitting of the model. In contrast, a model may capture the patterns as well as noise in the data and may perform poorly for unseen data which is considered as overfitting of the model [53]. The CV is one of the ways to prevent underfitting and overfitting of the learning models. Some of the important CV techniques are as follows:

Train Test Split method: The dataset is partitioned into train and test data. The default ratio of partitioning is 70:30 where 70 percent of data is used to train the model and 30 percent is kept aside to test the model [52]. Sometimes, the train test split is modified by dividing the training data (70 percent of data) into training and validation data, and 30 percent is used to evaluate the test performance of the trained model. This is used for the hypertuning of parameters of a model.

Leave One Out Cross Validation (LOOCV): In LOOCV, one sample, as a test sample, is removed from all samples and the remaining are used for training the classifier [54,55]. The prediction of the test sample is saved, and this process is repeated for all the samples. LOOCV is desirable where the number of samples (patients) is less as it is an exhaustive and time consuming technique. Leave-p-out CV is a generalized version of LOOCV where p samples are removed from the whole set as test samples [54].

K Fold Validation: In k-fold CV (K is a user defined number), k-1 folds of samples out of k folds are selected to train the proposed ICNNM model and the remaining one fold of samples is used for the validation process [56]. This process is repeated k-times for training and validating all folds. Sometimes, the K-fold CV is named based on the values of the K. For example, if K=10, then it is called as 10-fold CV technique.

Stratified K Fold Validation: In stratified K-fold, the samples are chosen from each class for each fold unlike in k-fold where the samples are selected randomly [57]. The

stratified K-fold is suitable for imbalanced data like cancer. The imbalanced datasets are those containing an unequal number of samples in different classes and their ratio is greater than a threshold [58].

1.4 Literature Review

The thesis contributes in developing computational methods to analyze the miRNA expression data, generated from biochemical methods, in cancer. The association of miRNAs in cancer diagnosis and treatment can be divided into two categories, biological or biochemical experiments (Section 1.4.1), and computational investigations (Section 1.4.2).

1.4.1 Biochemical Methods

Various biochemical investigations are conducted for identifying miRNAs in various types of cancer and in drug resistance [37, 59–62]. Some of the related studies are discussed in this section, and these are categorized into three types: i) identification of miRNAs in individual cancer, ii) identification of drug resistant miRNAs, and iii) identification of miRNAs in various cancer classes using pan-cancer data.

1.4.1.1 Identification of miRNAs in Individual Cancer

One of the earlier investigations that established the role of miR-15 & miR-16 in chronic lymphocytic leukemia (CLL) was in 2002 [14]. Thereafter, other types of cancers are also analyzed using miRNAs, and those are found expressed differently in normal and cancer patients [6]. Lu et al. [63] analyzed miRNAs from 334 leukemia patients and reported that miRNAs possess discriminating capabilities. MiRNA expressions of 363 patients from six cancers and 177 normal persons are analyzed in [64]. It is reported that 21 miRNAs are downregulated and 36 are upregulated out of a total of 228 miRNAs. In [65], the role of miRNAs in breast cancer is investigated. MiR-21 and miR-146 are found deregulated and revealed as key biomarkers in breast cancer diagnosis. While miR-21 is established as an oncogene and upregulated, miR-146 is found as a tumor suppressor. MiR-15, miR-16, miR-143, and miR-145, are reported as downregulated in abdominal cancer and responsible for tumor development [66]. These are also found resistant to multiple drugs for stomach cancer. In [67], deregulation of miRNAs is found to be one of the reasons for cancer initiation and progression in head and neck. Alpha protein is considered as a key biomarker for diagnosis of hepatocellular carcinoma (HCC), but its false negative rate is approximately 40% at early stages [68]. This is addressed by finding expression levels of miRNAs (miR-101, miR-32, and miR-96) in

HCC. Petillo et al. investigated the levels of expression in kidney subtypes and reported that shifts in expression levels of miRNAs may develop different tumor subtypes. For example, miR-424 and miR-203 are upregulated in papillary whereas downregulation of miR-203 is observed in benign oncocytoma [69]. In benign oncocytoma, its expression is downregulated as compared to normal tissues in the kidney. Besides the discussed cancers, there are multiple types of cancers like acute myeloid, brain, bone and soft tissue, cervical, colon, esophageal, lung, lymphoma, nasopharyngeal, rectal, pancreatic, prostate, skin, testicular, thyroid, uterine, and uveal. These cancers are investigated as a part of many international projects such as Gene Expression Omnibus (GEO), the Cancer Genome Atlas (TCGA), and the International Genome Cancer Consortium (ICGC).

1.4.1.2 Identification of Drug Resistant miRNAs

In the last two decades, the role of miRNAs in drug resistance of patients is evaluated in different types of cancers such as breast, colon, ovarian, lung, etc. One of the earlier investigations related to miRNAs and their role in drug resistance is performed by Miller et al. [70]. They reported that miR-221 and miR-222 control tamoxifen drug resistance in breast cancer. Further, clustering of miRNAs revealed eight and seven significantly overexpressed and underexpressed miRNAs, respectively. Pichiorri et al. [71] investigated vivo nucleolin (NCL), a major nucleolar protein that regulates the expression of a specific subset of miRNAs, which may cause breast cancer. Inhibition of NCL using guanosine-rich aptamers reduces the levels of miRNAs that are dependent on NCL and their target genes. Mencia et al. [72] studied the role of miRNAs associated with drug resistance in human colon cancer cells, treated with methotrexate (MTX), and reported the significance of miR-224 using the GeneSpring GX11.5 software. Kurokawa et al. [73] reported that miR-19b and its target mRNAs are up-regulated in response to 5-Fluorouracil and are thus responsible for the resistance to the drug treatment. Kumar et al. [74] applied locked nucleic acid (LNA) technology to investigate the role of miRNAs associated with drug resistance in an ovarian cancer cell line, (A2780/CP70) treated with cisplatin. The results revealed that five miRNAs are up-regulated in the cell line and six miRNAs are down-regulated as compared to cisplatin sensitive A2780 cells. Kitamura et al. [75] reported that miR-134, miR-487b, and miR-655 regulate drug resistance towards Gefitinib by targeting the gene MAGI2 in lung adenocarcinoma cells. These miRNAs regulated the epithelial-mesenchymal transition and then acquired resistance to the epidermal growth factor receptor tyrosine kinase inhibitor (EGFR-TKI). Hummel et al. [76] studied the role of miRNAs associated with drug resistance in esophageal adenocarcinoma (EAC) and squamous cell carcinoma (ESCC) using the drugs cisplatin and 5-fluorouracil. Chemotherapy-resistant sublines presented different miRNA signatures and responses to cisplatin vs 5-FU resistant cells from the cell line of the same tumor. Schotte et al. [77] discussed how miRNAs characterize genetic diversity

and drug resistance in pediatric acute lymphoblastic leukemia. Expression levels of 397 miRNAs are measured by quantitative RT-PCR in 17 control patients and 81 cases of pediatric leukemia. Resistance to the drugs vincristine and daunorubicin is shown by up-regulation of miR-125b, miR-99a, and miR-100 by approximately 20-fold [77]. Further, drug resistance in some other cancers like prostate [78], gastric cancer [79], human laryngeal cancer [80], and myeloma [81], is also explored. Apart from experimental studies of individual cancers, some review articles [82–86] also exist that provide deep insight into the association of miRNAs with drugs in cancer.

1.4.1.3 Identification of miRNAs in Various Cancer Classes using Pan-cancer Data

In this section, we discuss the studies that investigated the role of miRNAs in various cancers through pan-cancer analysis. For pan-cancer analysis, many public repositories such as 'The Cancer Genome Atlas (TCGA)' [1], 'Broad GDAC Firehose' [87], and 'Xena' [88] are available. In [89], the miR-sequence of 575 (normal and tumor) patients for 11 cancer types is extracted from the cancer genome atlas (TCGA) database. Two 'R' language packages, DeSeq2 and edgeR are used for finding differentially expressed miRNAs (DEmiRNAs). Twenty-one miRNAs out of 81 DEmiRNAs are found to be significantly expressed. Among those, 9 are found to be upregulated and 12 are down-regulated. Wong et al. [90] presented an online tool named OncomiR for identifying DEmiRNAs in various cancers. In this study, they integrated miR-sequence, RNA sequences, and clinical data from TCGA for identifying miRNAs responsible for tumor development, staging, grading, and survival analysis of cancer patients. They used ANOVA for staging and grading of cancer patients, and univariate Cox proportional hazard analysis along with Student t-test for survival analysis. In [91], the survival rate of European American (EA) and African American (AA) patients for 42 cancer types is studied. It is found that AA patients have a higher fatality rate than EA. Expression levels of miRNAs specific to AA patients are correlated with resistance to cancer treatments. Elevated levels of miRNAs (miR-15a, miR-17, miR-130-3p, miR-181a) are observed consistently among AA patients with thyroid, prostate, kidney, and breast. Sabbaghian et al. [92] analyzed miRNA expressions of 11 cancer types and established that the expression levels of miRNAs are different in healthy and cancer patients. They found 7 significantly DEmiRNAs whose expression levels were dysregulated in all cancers. Wu et al. [93] developed an early cancer detection study by analyzing miRNA expression profiles of approximately 15000 patients for 13 cancer types. They reported 100 cell-free immune-related miRNA profiles that showed fluctuations in levels in healthy and tumor patients.

1.4.2 Computational Investigations

The expression data for analysis may be labeled or unlabeled. Labeled data can be handled with supervised learning, where the labels help the model to classify the patients. Unlabeled data can be analyzed with unsupervised learning, such as clustering methods. Based on the above two learning techniques, miRNA expression data can be analyzed for patient classification and miRNA selection.

1.4.2.1 Identification of miRNAs in Individual Cancer

Here, some techniques that are very specific to miRNA or gene selection using expression data are discussed. Pal et al. [94] developed a feature grouping and selection method based on fuzzy mutual information (FMIMS) and SVM classifier to select miRNAs using their expression profiles. The SVM classifier is used to generate the miRNAs groups and fuzzy mutual information helps in selecting the best group of miRNAs. The more the fuzzy mutual information of the group, the higher the group's significance. An improvement of this work is presented in [43] where a fitness function for particle swarm optimization is developed. Navon et al. [95] developed a fold change ranking method to find characteristics of miRNAs in multiple cancers. They used data consisting of paired samples from normal and tumor origins. The benign and malignant tumors are extracted from the same patient. The miRNAs that are dysregulated in eight cancer datasets are considered important ones. Piao et al. [96] developed an ensemble learning algorithm for miRNA selection based on a feature subset technique to classify multi-class cancer patients. Leidinger et al. [34] made a hypothesis that miRNAs showing fold change values greater than 2 are considered deregulated and selected 51 miRNAs using that hypothesis. Sathipati et al. [97] combined an inheritable bi-objective combinatorial genetic algorithm and SVM classifier to select miRNAs in breast cancer.

Some popular gene selection techniques can also be used for miRNA selection. Peng et al. [33] presented a feature filtering method based on minimum redundancy-maximum relevance (MRMR) criteria. The MRMR features are obtained through first-order incremental learning. Selected features are evaluated with support vector machine (SVM), naive Bayes (NB), and Linear Discriminant Analysis (LDA) classifiers. Leave-one-out cross-validation (LOOCV) method is used in the process. In [38], genes are recursively eliminated by checking the performance of the SVM classifier (SVMRFE). The set of genes is considered important which provides the highest classification accuracy. Mundra and Rajapakse [98] developed an integrated approach using the support vector machine-based recursive feature elimination (SVMRFE) with MRMR approaches. The method is computationally more expensive than both SVMRFE and MRMR. Alok et al. [99] developed a null linear discriminant analysis-based algorithm for gene selection. Selected genes are validated using a nearest neighbor classifier after gene selection. Ghalswal et al. [100] developed a gene selection approach to find out the jointly discriminative group

of genes. Guo et al. [101] used the gradient descent approach with an objective function involving regularized logistic regression in order to obtain the global optimum. A graph-based gene selection model with minimum redundancy and maximum relevance is presented in [102]. The method uses social network algorithms, such as the maximum weighted complete sub-graph of a graph to maximize the relevance of genes in a particular class. Edge centrality measure is also used to deal with the local and global structure of the graph to rank the genes. A hybrid ensemble based fuzzy approach is developed in [103] for gene selection using expression data. An ensemble of feature selection methods is used to create a gene pool for cancer patient classification. In [104], a score is developed by a weighted combination of mutual information, f-classification, and chi-squared test results to rank genes. Maji and Pal [36] presented a fuzzy rough set based information measure (FRMI). Ensemble learning using fuzzy rough approach is developed in [105] for gene selection using expression data. First, informative features are selected using active learning based on a fuzzy rough set, and then an ensemble of classifiers is used for patient classification. A fuzzy similarity relation by using feature neighborhood and sample class information is developed in [106] for selecting important genes in cancer patient classification.

1.4.2.2 Identification of Drug Resistant miRNAs

Several investigations deal with the task of identifying miRNAs related to drug resistance [37, 59–62]. In [37], a set based approach, called SPEM, is introduced where a set S is extended to S^+ utilizing the z number concept to handle the uncertainty in miRNA selection with confidence. Each element (miRNA) of the set is presented as a tuple, where the first part of the tuple is the relevance of the element to a particular category (sensitive or drug resistant), and the other part determines the confidence of selecting the element where the confidence is computed using histogram-based granular probability. A data integration method for protein-protein, miRNA-mRNA, and miRNA-lncRNAs interaction is developed by Yang et al. [60]. A graphical network of biological molecules is constructed and random walk with restart is used to find a probability that prioritizes the lncRNAs and miRNAs related to chemoresistance. A ROC curve using the results of the leave-one-out cross-validation procedure is plotted, and the Wilcoxon sum rank test is used to validate the identified top lncRNAs in 60 cancer lines. Moughari and Eslahchi [61] presented a classification of drug sensitivity method using machine learning (CDSML) based on manifold learning techniques. The method is focused on extracting binary responses from various inputs like gene expression, copy number variation, and mutation profiles. In the process, a combination of standardization and normalization of inputs is performed. K-nearest neighbour is used to impute missing values in the dataset. The classification performance of the CDSML is compared with different classification methods and regression algorithms. The drug sensitive dataset is collected from the 'Genomics of Drug Sensitivity in Cancer' (GDSC) database. Xin et al. [59] examined the

role of miRNAs using fulvestrant-resistant MCF7-FR cells in breast cancer and compared the global miRNA and mRNA expression patterns. Fourteen downregulated miRNAs are identified in MCF7-FR cells. Computational tools like TargetScan and PITA are used to predict potential target genes, from which a negative correlation is found between the expression of these miRNAs and their predicted target mRNA transcripts. The results suggested a significant role of miRNA-regulated gene expression in the onset of breast cancer antiestrogen resistance.

In [107], four graph convolutional networks utilizing unique combinations of features, including drug fingerprints, drug label encoding, miRNA expression profiles, and miRNA GO-based similarities, are constructed, and then the latent representations of drugs and miRNAs from the sub-networks are learnt. Attentive representations of drugs and miRNAs from their corresponding latent representations are derived using an attention neural network. A dot product of the attentive representations is computed to predict the miR-drug resistance association. Zheng et al. [108] integrated a neural architecture search and a graph isomorphism network to predict miR-drug resistance association by leveraging features instead of graph interactions. They mapped the miRNA sequences into miRNA k-mer sparse matrices and then utilized the non-negative matrix factorization technique to represent the k-mer sparse matrix into a low-dimensional space. Later, they integrated various graph isomorphism networks to construct descriptors as features to find suitable association models [109]. In [110], a fusion of multi-view contrastive learning and graph collaborative filtering (GCF) approach is utilized to determine the association of miRNAs and drugs, where the model considers the miR-drug topology as a bipartite graph and exploits GCF to find homogeneous neighborhood features representation from the layers of the graph neural network. Guan et al. [111] developed a feature-integration method using a CNN and a deep neural network to determine the association between miRNAs and drugs. Drug miRNA association is treated as a bipartite graph and then a structural embedding is used to find the topological properties of the graph. The Word2vec technique is employed to develop attribute features of drugs and miRNAs. Finally, these two types of features are utilized as input to CNN and DNN to combine features and predict the target miRNAs of drugs. Yu et al. [112] developed a web server to determine the impact of drugs on miRNAs. They employed k-mer, sequence data, and MACCS fingerprints to characterize the miRNAs and drugs. The modulation of miRNA expression by the drugs was subsequently predicted using random forests. In [113], a model to determine the miR-small molecule association is developed utilizing non-negative matrix factorization (NNMF) and regularized least square (RLS) techniques. Initially, the integrated similarity matrices are interpolated by utilizing symmetric non-negative matrix factorization (SymNMF), subsequently, the Kronecker products of the newly integrated similarity matrices are computed, and ultimately derived a function for calculating the association probabilities of disease-miRNA pairs via regularized least squares (RLS). In [114], an interpretable deep learning framework, comprising a dual-channel representation technique and a heterogeneous graph

global-attention network, is developed to predict miRNA-drug sensitivity associations.

1.4.2.3 Identification of miRNAs in Various Cancer Classes using Pan-cancer Data

Pan-cancer miRNA expression data deals with multiple cancer classes. In [115], an ensemble of K-nearest neighbor and decision tree is used to classify patients from 42 cancer classes using 64 miRNAs. Lopez et al. [116] developed a miRNA selection algorithm by combining ensemble classifiers and feature selection methods such as gradient boosting, random forest, and logistic regression. They selected 100 miRNAs to classify patients in each cancer using an ensemble of 8 classifiers. In [117], a framework, called Cancer*Sig*, is introduced to find miRNAs responsible for classifying 15 types of cancer. Cancer*Sig* uses an inheritable bi-objective combinatorial genetic algorithm to optimize the SVM classifier performance and number of miRNAs. In [118], three feature selection techniques, lasso, SVM-RFE, and random forest, are used in the ensemble for miRNA selection. The selected miRNAs are used to classify patients using an ANN as a classifier. The findings are further validated using quantitative polymerase chain reaction (qPCR) experiments. Cheerla and Gevaert [119] classified patients from 21 cancers using the radial basis function as a metric in the support vector machine. They used miRNA pan-cancer expression from the Cancer Genome Atlas (TCGA) database. Li et al. [120] combined gene algorithms and K-nearest neighbor to classify 31 types of cancer using gene expression data from the TCGA database. An ensemble of 14 classifiers, with 182 combinations using the R package 'glm', is presented for pan-cancer classification in [121]. Ruidong et al. [122] developed a database named as CancerMIRNome to analyze and visualize miRNA expression data of the 10554 patients from 33 cancer types and 28633 samples from 32 cancer types. The database is useful for differential expression analysis, survival analysis, and functional enrichment analysis. It also involves identifying miRNAs using DE analysis using LASSO, dimensionality reduction techniques, and univariate survival analysis. In [123], a prognosis prediction method using univariate and multivariate Cox regression techniques for cancer-specific survival (CSS) using miRNA expression data is presented. The Univariate and multivariate Cox regression methods helped in identifying a five miRNA set linked to CSS of breast cancer patients. Kaplan–Meier (KM) and receiver operating characteristic (ROC) curves were also used to validate the clinical relevance of the five-miRNA signature. Shuting et al. [124] introduced a framework for patients' survival in various cancer types using miRNA expression data. Firstly, survival analysis is conducted on cancer-stratified and drug-stratified patient subpopulations to predict the miRNAs that may function as drug-specific survival biomarkers. Secondly, miRNAs showing significant correlation with survival outcomes of patients are identified. In [125], a pan-cancer analysis is performed to identify miRNA-gene associations by integrating multi-omics datasets, including DNA methylation, copy number aberrations, miRNA, and gene expression data, altered in specific DNA regions using a Lasso-based

regression technique. The analysis is performed on 7294 sample datasets of 18 cancer types from the TCGA database. Kuthethur et al. [126] presented a framework for identifying deregulated miRNA during breast cancer progression and related to homologous recombination deficiency in breast cancer patients. In [127], a framework is introduced for identifying miRNAs that can suppress the various types of tumors by analyzing pan-cancer expression data of 14 tumor types with the help of the 'Limma' package (a R language package).

Recently, deep convolutional neural networks became popular in bioinformatics, natural language models, and biomedical health informatics for their ability to deal with HDC data. Lopez et al. [128] meta-heuristically optimized the hyperparameters of a CNN model using genetic algorithms, where the number of hidden layers is already fixed at 3. They classified 29 types of cancers using miRNA expression data with no mention of normal samples. Li et al. [129] developed an integrated approach where features are first selected using the elastic net algorithm. Those features are then used to classify multiclass cancer patients using a deep neural network. In [130], a two-dimensional CNN model is provided to classify pan-cancer samples using images as input data. Assuming nearby genes interact better, gene expression data is arranged based on chromosome loci, and gene expression of a sample of shape 10302×1 is reshaped into a 102*102-sized image. In this work, 33 cancer types are classified without considering the gene expressions of normal samples. This issue is resolved by Mostavi et al. [131] by considering both cancer and normal gene expression data and classifying them using three different CNN models 1D CNN, 2D Vanilla-CNN, and 2D Hybrid-CNN. In [132], genes are first selected using the Laplacian score, and then the selected genes are fed to a 6-layer 1D CNN model for patient classification. Mohammed et al. [53] used LASSO [133] for gene selection and then used a stack of 5 1D CNN models, with one layer in each CNN, in an ensemble framework. For layer ensemble, the K-nearest neighbor algorithm is used. Raghu et al. [134] developed a miRNA selection method to detect tissue of origin (TOO) for metastatic cancer using various classifiers such as decision tree, random forest, logistic regression, and deep neural networks.

1.5 Motivation

It is well established in the literature (See Section 1.4) that miRNAs play an important role in cancer diagnosis and treatment. It is also evident that not all miRNAs are responsible for cancer and identification of responsible miRNAs in an unknown patient is a crucial task from a computational perspective. Therefore, the aim of the thesis is to identify a subset of miRNAs for two-class drug resistant expression data and also multi-class pan-cancer expression data, and thereafter using those miRNAs for the classification of cancer patients. The summary of the motivations for each chapter is provided as follows:

1. To select a set of drug resistant miRNAs, from two class drug resistant expression data, by utilizing the known drug resistant miRNAs from existing literature as biological knowledge.

2. To identify a set of miRNAs from drug resistant expression data that efficiently classifies the control and drug resistant patients by using information about known patient labels.

3. To find relevant miRNAs for multiple cancer classes in pan-cancer expression data and to improve the accuracy in classifying patients for various classes.

4. To identify relevant miRNAs and to classify patients for various classes in pan-cancer expression data by enhancing the methods developed in the previous chapter.

Now we discuss the motivation for developing various techniques in each chapter in detail. The four contributory chapters are 2, 3, 4, and 5. In Chapter 2, the motivation is to identify unknown drug resistant miRNAs by using a known set of drug resistant miRNAs. This is achieved by developing feature selection methods, namely, Euclidean distance with weighted fold change (EDWFC) and histogram-based clustering and Euclidean distance with fold change-based ranking (HCEDFCR). EDWFC and HCEDFCR are feature selection methods for selecting drug resistant miRNAs. While EDWFC is proposed to select user defined number of drug resistant miRNAs by exploring the non-linear relationships among various miRNAs using the biological knowledge of known drug resistant miRNAs, HCEDFCR is proposed to automatically select a group of drug resistant miRNAs using histogram based clustering technique. The strength of EDWFC lies in utilizing biological knowledge to find a power of fold change which minimizes the average rank of known drug resistant miRNAs. As EDWFC is a product of the Euclidean distance and fold change with varying power, it helps to explore the search space using two different similarity measures based on the distance and ratio of expression values. Further, multiplying two linear functions results in a quadratic function and applying weight as a power increases its nonlinearity beyond the quadratic relation, which helps in identifying non-linear relationships, known as well as unknown drug resistant miRNA. The strength of HCEDFCR lies in identifying automated groups of miRNAs instead of user specified choice of the number of miRNAs using a histogram-based grouping method. Groups of useful miRNAs are required if the user has not determined about how many miRNAs will be enough for classifying an unknown (drug resistant or not) patient. Further, the ranking of groups using HCEDFCR helps to find the most important ones. Note that the product of the Euclidean distance and the fold change value, using the control and resistant miRNA expressions, is used to rank the clusters and the miRNAs inside the clusters. The developed methods are efficient for both small sample size and large sample size datasets as these are developed using classical measures or techniques such as Euclidean distance, fold change, and histogram.

In Chapter 3, the motivation is to select a set of miRNAs for which the classification accuracy in distinguishing control and drug-resistant patients is maximized. A miRNA selection method called a weighted framework for integrating fuzzy rough set-based relevance and redundancy entropies (WFIFRRRE) is introduced. The weights in WFIFRRRE, assigned to relevance and redundancy entropies, are varied in a supervised manner to maximize the F score in the patient classification process. Note that, while in Chapter 2, information about known drug resistant miRNAs is utilized, here in Chapter 3, information about known patients is utilized. The strength of WFIFRRRE lies in integration of relevance and redundancy entropies those help in finding a set of miRNAs that provides high classification accuracy. Moreover, the judicious integration of fuzzy set and rough set not only helps in handling uncertainty in control and drug-resistant class overlapping but also helps in determining the exactness of class size. A tradeoff between relevant and redundant miRNAs for drug resistant class is also explored. The method is suitable for both small and large datasets as fuzzy rough set are utilized to develop them.

In Chapters 4 and 5, the motivation is to select miRNAs in multiple types of cancer and to classify patients through pan-cancer analysis. In Chapter 4, an interpretable convolutional neural network model, called ICNNM, is developed. While the CNN model is developed for classifying patients from different cancer classes, the interpretable approach is developed using SHapley Additive exPlanations (SHAP) values for selecting miRNAs in various cancer classes. Higher SHAP values are obtained for miRNAs those help in accurate prediction of the cancer class of a patient. The strength of the developed method lies in optimizing the hyperparameters of a one-dimensional CNN model, in single objective framework, which helps in maximizing the classification accuracy of patients. The strength of SHAP values based interpretation lies in selecting those miRNAs which contributes in the classification process. The method is suitable for large sample datasets. For small sample datasets, CNNs may suffer from an overfitting issue.

The method developed in Chapter 4 is enhanced in Chapter 5 to further improve the classification accuracy of patients and to select the relevant miRNAs for multiple cancer classes. This is achieved by developing a multi-objective framework for optimizing hyperparameters of a 1D CNN (MOHCNN) for patient classification and a set-theoretic explainable AI-based attribution score (STEAAS) for miRNA selection, using pan-cancer expression data. The strength of MOHCNN lies in utilizing training error, validation error, and the number of training parameters as objective functions, and using Bayesian optimization with the tree Parzen estimator for optimizing the hyperparameters in the patient classification process. The strength of STEAAS lies in integrating two explainable methods, Shapley additive explanations (SHAP) and Bayesian Local Interpretation model agnostic explanation (BayLIME) for ranking and selection of miRNAs for various cancer classes. Integration of two explainable methods provides more comprehensive and robust assessment of data relationships than any single method. Here also the methods

are suitable for large sample datasets.

1.6 Scope and Organization of the Thesis

Considering this, two major problems, the identification of miRNAs associated with drug resistance using expression data and the detection of the miRNAs that efficiently classify various cancers using pan-cancer expression data, are addressed in this thesis. In this regard, four contributory chapters, namely, Euclidean distance with weighted fold change based score for identifying drug resistant miRNAs, integrating fuzzy rough set-based entropies for identifying drug resistant miRNAs, identifying pan-cancer and cancer subtype miRNAs using interpretable convolutional neural network, and multi-Objective framework for optimizing CNN and set-theoretic explainable AI based attribution score for selecting miRNAs in pan-cancer data, are presented to address the mentioned problems. A schematic diagram representing the organization of the thesis is provided in Fig. 1.4. The various datasets are collected from public data repositories such as Gene Expression Omnibus (GEO) [135] and the Cancer Genome Atlas (TCGA) [136]. The datasets available on these repositories are deposited by various research groups after taking care of ethical issues. For example, the miRNA expression data for various cancer types are submitted to TCGA by the Michael Smith Genome Center, Canada, where the ethical issues are handled. The organization and scope of the chapters are provided in the next sections.

1.6.1 Selecting Drug-Resistant miRNAs in Cancer using Euclidean Distance with Fold Change based Score

In Chapter 2, two computational methods, called 'Euclidean distance and weighted fold change based ranking' (EDWFC) and 'Histogram based clustering and the Euclidean distance with fold change based ranking' (HCEDFCR), are presented to identify drug resistant miRNAs in cancer [137]. In EDWFC, the Euclidean distance and fold change, computed using the averages of control and resistant expressions, are multiplied by varying the power of fold change to minimize the average rank of biologically known miRNAs. The miRNAs are then sorted in descending order according to the EDWFC values to select the top miRNAs. In HCEDFCR, the miRNAs are first divided into different clusters using a novel histogram-based clustering method. The product of the Euclidean distance and the fold change value, using the control and resistant miRNA expressions, is then used to rank the clusters and the miRNAs inside the clusters. Finally, the top ranked miRNAs from the top ranked clusters are considered as an important group of identified miRNAs.

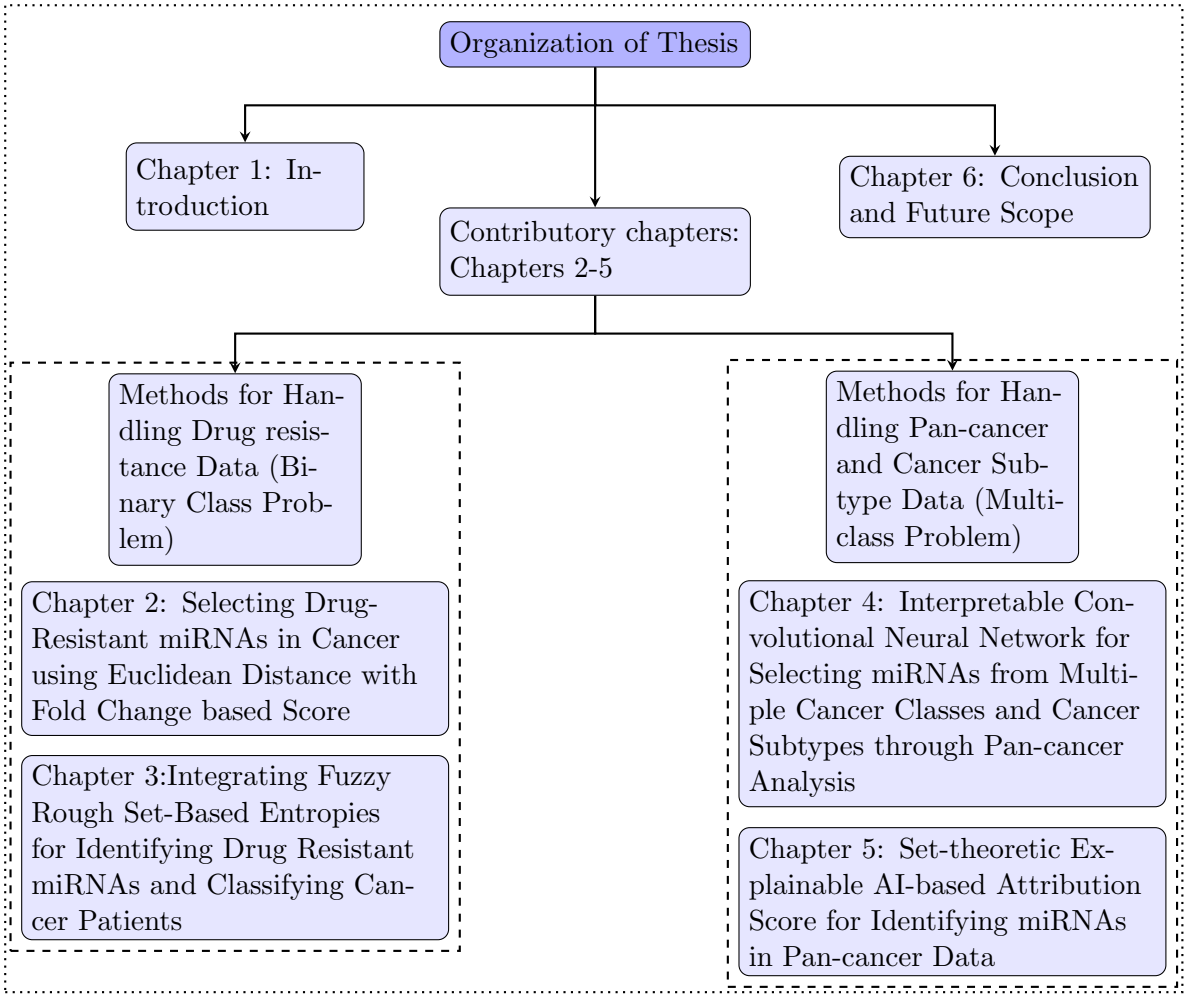


Figure 1.4: Structural organization of thesis and the grouping of proposed contributions.

The performance of the developed methods is evaluated on eight datasets consisting of miRNA expressions of control and drug resistant patients. Self-evaluation of the developed methods and their comparison with related methods are discussed. For self-evaluation of EDWFC, the power of fold change at which the average rank of known miRNAs is minimized for various datasets is reported. EDWFC is compared with related methods, used for miRNA and gene selection, such as SPEM [37], 1D-CNN [131], CSPS [138], MRMR [33], SVMRFE [38], SVMRFE with MRMR [98], FSNLDA [99] and Lasso [139]. A percentage of top ranked miRNAs is selected for each method and used for patient classification using SVM [140] and RF [140] classifiers. The EDWFC performed better in most of the cases for different datasets. For HCEDFCR, the ranks of clusters and ranks of miRNAs inside each cluster are reported to show the efficiency of the model. HCEDFCR is also compared with related clustering methods. The cluster ranks and ranks of miRNAs, obtained using HCEDFCR, are found to be superior to the related methods.

While EDWFC uses biologically known miRNAs to select top miRNAs for patient

classification, HCEDFCR uses clustering techniques for miRNA selection without any biological knowledge. In the subsequent chapters, the selection of miRNAs is performed based on the patient classification performance without using knowledge about any known miRNA.

1.6.2 Integrating Fuzzy Rough Set-Based Entropies for Identifying Drug Resistant miRNAs and Classifying Cancer Patients

In Chapter 3, two new z score based fuzzy rough relevance and redundancy entropies are developed and then a weighted framework is introduced to integrate the entropies for ranking and selecting miRNAs for classifying control and drug resistant patients. Here, two key components of soft computing, fuzzy set and rough set are utilized. The methodology is called as weighted framework for integrating fuzzy rough set-based relevance and redundancy entropies (WFIFRRRE). The z score is used to compute the fuzzy membership of expression values required for both entropies. Fuzziness deals with the overlapping nature of miRNA expression profiles and rough set helps in determining the exact class size. The weights in WFIFRRRE, assigned to relevance and redundancy entropies, are determined in a supervised manner to maximize the F score used for validating the classification performance in discriminating between the control and drug resistant patients.

The performance of the miRNAs selected by WFIFRRRE is evaluated using binary SVM [140], random forest [140], and Naive Bayes [140] classifiers on eight drug resistant cancer datasets. Binary SVM, NB, and RF are used for classifying control and drug-resistant patients because there are only two classes present in the drug resistance miRNA expression data. In our investigation, the number of samples is low, and the boosted trees may suffer from overfitting issues [141]. The performance of WFIFRRRE is also compared with some well-known miRNA and gene selection techniques such as EDWFC [137], SPEM [37], MRMR [33], SVMRFE [38], SVMRFE-MRMR [142], FRMI [36], CBFS [143], FSNLDA [144], FSHDD [145], GSPSO [146], BPSOFS [147] and Lasso [139]. It is observed that WFIFRRRE performed better than the related methods in most of the cases. The classification accuracy values, in terms of the F1-scores, achieved at different percentages of miRNAs shows that it decreases for higher percentages of selected miRNAs. From the results it is clear that not all miRNAs are responsible for drug resistance and it signifies the importance of WFIFRRRE.

Chapters 2 and 3 deal with the methodologies developed for the detection of miRNAs those help in classifying control and drug resistance patients using expression data. The next chapters (Chapters 4 and 5) deal with the identification of miRNAs in various types of cancer using pan-cancer miRNA expression data.

1.6.3 Interpretable Convolutional Neural Network for Selecting miRNAs from Multiple Cancer Classes and Cancer Subtypes through Pan-cancer Analysis

As discussed earlier, pan-cancer miRNA expression data is high dimensional and complex (HDC) in nature and convolutional neural networks are proven to be good performer due to their ability of finding patterns in complex data. Therefore, an interpretable convolutional neural network model (ICNNM) is developed for classifying patients and selecting miRNAs for various cancer classes. The ICNNM is a one dimensional CNN model. The layers and other hyperparameters are optimized using Bayesian optimization with multivariate tree parzen estimator (BoMTPE). An interpretable approach is developed using SHapley Additive exPlanations (SHAP) values for explaining the behavior of ICNNM. This approach helps in introducing an attribution score for identifying relevant miRNAs using SHAP values. The attribution scores are assigned higher values for those miRNAs which help in the accurate prediction of tumor class of patients by utilizing the game theory concept in computing the SHAP values.

The model is evaluated on seven datasets among which six datasets are derived from a single TCGA pan-cancer dataset [88], and the breast subtype data is downloaded from the study in [8]. The ICNNM is seen to perform better as compared to related techniques such as three variations of the CNN model (Base-CNN [131], Ens-CNN [53], and LS-CNN [132]), random forest [140], Gboost [140], XGboost [140], Catboost [148], and SVM [119]. The performance is evaluated in terms of F-score, discriminability power of expressions between normal and cancer classes, and biological significance of the selected miRNAs. The performance of ICNNM in terms of accuracy and F-score varies from 0.98 to 0.99 and 0.91 to 0.99, respectively. Some top miRNAs, achieving high attribution scores by ICNNM, are found to be the key biomarkers in various cancers.

In ICNNM, the 1D CNN is optimized in a single objective framework for classifying patients and SHAP is used for global interpretation and selection of miRNAs. The miRNAs in various cancers are identified based on the attribution score obtained using normalized SHAP values. In the next chapter, a 1D CNN is optimized in multi-objective framework, and a set-theoretic explainable AI-based attribution score is developed for the selection of miRNAs in different cancers. The relevance of miRNAs, as well as their reliability in each class are obtained using the score.

1.6.4 Set-theoretic Explainable AI-based Attribution Score for Identifying miRNAs in Pan-cancer Data

In Chapter 5, a multi-objective framework for optimizing hyperparameters of a 1D CNN, called MOHCNN, and a set-theoretic explainable AI-based attribution scores (STEAAS)

for miRNA selection are developed. In MOHCNN, the number of layers and other hyperparameters of the convolutional neural network is optimized using Bayesian optimization with tree parzen estimator in a multi-objective framework. Three objective functions, training error, validation error, and a number of training parameters, are considered for the optimization of hyperparameters. The training and validation error represent the model's learning capability, and the number of training parameters denotes the computational complexity of the model. A new set-theoretic explainable AI-based score is also defined, which uses the $Z-number$ concept, to find the importance of a miRNA and the level of reliability for that miRNA belonging to a particular cancer class. In STEAAS, the attribution score is represented as an ordered pair where the first value represents the importance of the miRNA computed using a class score in a particular cancer class, and the second part denotes the reliability score of the miRNA belonging to that class. The class score of a miRNA is computed by considering the average of BayLIME (Bayesian Local Interpretable Model Agnostic Explanations) [149] and SHAP (Shapley Additive exPlainations) [150] values of the patients belonging to that class. The reliability score utilizes the concept of Gini index for measuring the spread of expression values for all patients of a miRNA in a class. Finally, the class score and the reliability score of each miRNA for each cancer class are combined in a set-theoretic manner as an attribution score for selecting the relevant miRNAs.

The performance of the developed model is evaluated on 7 datasets. The developed MOHCNN model is compared with 4 relevant CNN models (ICNNM [151], Base-CNN [131], Ens-CNN [53], and SixL-CNN [132]) and 3 miRNA selection methods (EFS [116], SVM-RBF [119], and Onco-Cls [121]) in pan-cancer. Further, MOHCNN is also compared with two highly cited boosted classifiers, such as extreme gradient boosting (XGBoost) [152] and Catboost [148]. The proposed MOHCNN performed better than all the related CNN models used for comparison during training. The test performance of MOHCNN is also observed to be superior to the related methods in most cases. The performance of MOHCNN in terms of accuracy, F-score, and MCC ranges from 0.99 to 1.00, 0.97 to 1.00, and 0.96 to 1.0, respectively. The miRNAs selected by STEAAS are validated using OncomiR [90] and ENCORI/Starbase [153] database and related existing studies. Most of the miRNAs selected by STEAAS are found to be mentioned as key biomarkers in various studies. For example, hsa-miR-125a and hsa-miR-30a are selected from lung and breast subtype datasets and these are also mentioned as biomarkers in those cancers in [154] and [155], respectively.

1.6.5 Conclusion and Future Scope

Expression profiles of miRNAs are different in normal and cancer patients, and even in cancer patients expressions are different for those who are treated with drugs. Identifying the miRNAs responsible for accurate patient classification is one of the ways to increase

the survival rate among cancer patients. The role of miRNAs in drug resistance and various cancers is explored in the different chapters of the thesis. The conclusion and future scope of the developed methodologies are discussed in Chapter 6. Brief results, key findings, importance, and limitations of the various methods are provided. Some possible modifications of the developed techniques and their usage beyond miRNA detection are also mentioned.

Chapter 2

Selecting Drug-Resistant miRNAs in Cancer using Euclidean Distance with Fold Change based Score

2.1 Overview

The importance of developing computational methods for identifying miRNAs associated with drug resistance in cancer is discussed in Chapter 1. In analyzing miRNA expression data for drug resistance, the number of miRNAs is much greater than the number of samples (patients) which makes the identification of drug resistant miRNAs a difficult task. One more challenge is that not all miRNAs are responsible for drug resistance in patients. Hence, the objective is to identify the miRNAs those help in classifying control and drug resistant patients using expression data. While, the control group consists of patients who have not received chemotherapy, the resistant group consists of patients who have received drug treatment and developed drug resistance. To find the miRNAs associated with drug resistance in cancer, two computational methods using biological knowledge are discussed in this chapter. The methods are as follows:

1. The Euclidean distance and weighted fold change based ranking (EDWFC) [137]
2. Histogram based clustering and the Euclidean distance with fold change based ranking (HCEDFCR) [137]

In EDWFC, the Euclidean distance and fold change, computed using the averages of control and resistant expressions, are multiplied by varying the power of fold change

from 0.1 to 1.5 to find a value which minimizes the average rank of known miRNAs (obtained using biological knowledge) for each dataset. In HCEDFCR, the miRNAs are first divided into clusters using a novel histogram-based clustering method. The product of the Euclidean distance and the fold change value, using the control and resistant miRNA expressions, are then used to rank the clusters and the miRNAs inside the clusters. Finally, a portion of the miRNAs is selected from the top of the rank list in every cluster. The novelty of this work lies in i) utilizing the known drug resistant miRNAs from existing literature as biological knowledge to identify the relevant miRNAs, ii) formulating an interclass distance and weighted fold change based miRNA ranking method that minimizes the rank of known miRNAs, and iii) developing a histogram based clustering method by using valley concept.

Eight drug resistant miRNA expression datasets are used for the performance evaluation of the developed methods. The EDWFC is compared with related miRNA and gene selection algorithms, and HCEDFCR is compared with relevant clustering techniques. For both the methods, the selected miRNAs are compared with those obtained using various techniques. Further, the classification performance of the selected miRNAs by EDWFC is measured in terms of sensitivity, specificity, accuracy, F-score, and MCC using SVM and random forest (RF) classifiers. Leave one out cross-validation technique is used.

The rest of the chapter is organized as follows: The cancer datasets are summarized in Section 2.2. The schematic diagrams of EDWFC and HCEDFCR are provided in Section 2.3. The methodologies for EDWFC and HCEDFCR are described in Section 2.3.1 and 2.3.2, respectively. The results are reported in Section 2.4. The significance of the results is discussed in Section 2.6.

2.2 Datasets

Eight different cancer datasets used in this study are: breast cancer [71], colon cancer [73] treated with fluorouracil (Colon_FU), colon cancer [72] treated with methotrexate (Colon_M), esophageal cancer [76] treated with cisplatin (Esophageal_CIS), esophageal cancer [76] treated with fluorouracil (Esophageal_FU), lung cancer [156], lymphoblastic leukemia [77], and ovarian cancer [74]. These datasets are collected from the Gene Expression Omnibus (GEO), a publicly available data repository. The datasets for which the miRNAs associated with drug resistance are clearly defined in the related peer-reviewed journals, with available expressions for both the control and drug resistant groups, are selected.

The breast cancer dataset consists of 12 (6 control and 6 resistant) patients and 654 miRNA expressions [71]. While the colon cancer treated with fluorouracil dataset consists of 8 (4 control and 4 resistant) patients and 723 miRNA expressions [73], the colon

cancer treated with methotrexate dataset consists of 6 (3 control and 3 resistant) patients and 723 miRNA expressions [72]. Both of the esophageal cancer datasets consists of 12 (6 control and 6 resistant) patients and 847 miRNA expressions [76]. Further, the lung cancer dataset consists of 8 (4 control and 4 resistant) patients and 377 miRNA expressions, the lymphoblastic leukemia dataset consists of 58 (29 control and 29 resistant) patients and 365 miRNA expressions and the ovarian cancer dataset consists of 14 (7 control and 7 resistant) patients and 727 miRNA expressions. In [71], 4 miRNAs out of 654 are pointed out to be the responsible ones for drug resistance in breast cancer. Similarly in [73], 1 out of 723 miRNAs is identified as the most differentially expressed in colon cancer dataset treated with fluorouracil. In [72] also 1 miRNA is identified as the most differentially expressed in colon cancer treated with methotrexate. In [76], 4 miRNAs out of 847 are identified as the responsible ones in esophageal cancer treated with cisplatin and 3 miRNAs out of 847 are pointed out as the responsible ones for esophageal cancer treated with fluorouracil. In [157], 3 miRNAs out of 377 are identified as the most differentially expressed ones for lung cancer, 6 miRNAs out of 365 are identified as the responsible ones in lymphoblastic leukemia in [77] and finally 1 miRNA out of 727 is identified as the most differentially expressed one in ovarian cancer [74]. A summary of the used datasets is reported in Table 2.1.

Table 2.1: Summary of the datasets used.

Cancer Type	Total No. of miRNAs	No. of Control samples	No. of Resistant patients	No. of Identified miRNAs
Breast [71]	654	06	06	04
Colon_FU [73]	723	04	04	01
Colon_M [72]	723	03	03	01
Esophageal_CIS [76]	847	06	06	03
Esophageal_FU [76]	847	06	06	02
Lung [156]	377	04	04	03
Lymphoblastic [77]	365	29	29	06
Ovarian [74]	727	07	07	01

2.3 Developed Methods

In this section, first EDWFC and HCEDFCR are discussed briefly along with their schematic diagrams, and then these are described in detail. In EDWFC, the averages of control and resistant expressions are used for computing the Euclidean distance and fold change. The Euclidean distance is then multiplied with fold change and the power of fold change is varied such that the rank of the known miRNAs associated with drug resistance is minimized. Fig. 2.1 shows the block diagram of EDWFC method.

In HCEDFCR, a histogram is plotted using the expression values of the miRNAs of the resistant group. A valley concept is used to form the clusters, where a valley is represented by the bin whose left and right bins are higher. Then, the product of the

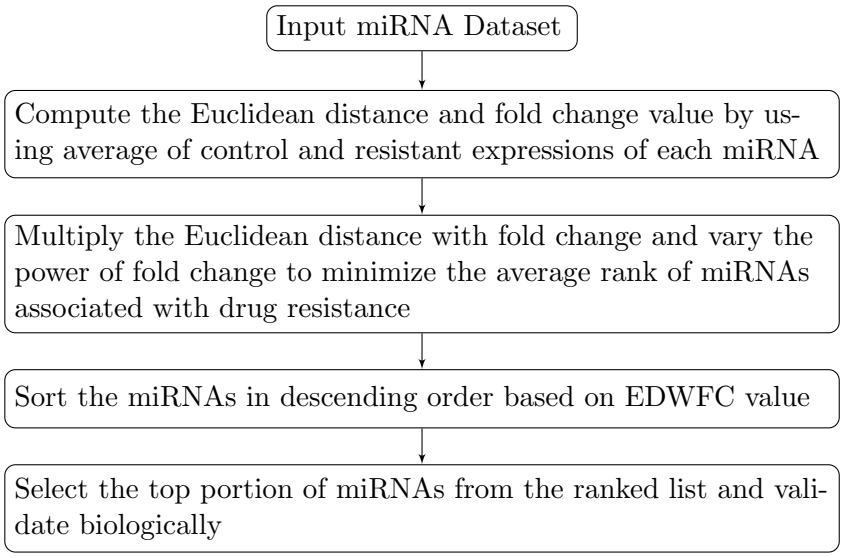


Figure 2.1: Schematic diagram of EDWFC method.

Euclidean distance and the fold change value, using the control and resistant miRNA expressions, is computed and used to rank the clusters and also the miRNAs inside the clusters. Fig. 2.2 shows the schematic diagram of HCEDFCR method.

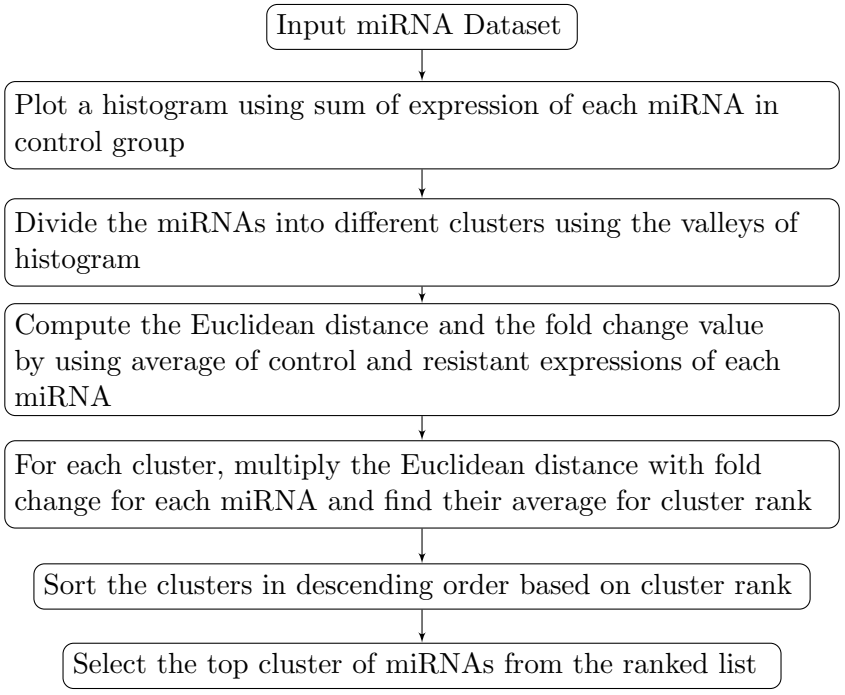


Figure 2.2: Schematic diagram of HCEDFCR method.

2.3.1 EDWFC

A ranked list of miRNAs is provided by the Euclidean distance with weighted fold change (EDWFC) method. The different steps of computing EDWFC are now discussed. These

are followed by the pseudocode in Algorithm 1. Given a set of miRNA expression values for the control and resistant group data for a particular type of cancer, the mean of the expression values for every miRNA is calculated for both groups. Thus, two mean values are obtained for each miRNA, one in the control data group and another in the drug resistance data group. This helps in identifying the centers of the classes. The Euclidean distance between these two means is then calculated for each miRNA. The aim is to find the separation between the control and drug resistant classes for a particular miRNA. Thereafter, fold change for each miRNA is computed and the Euclidean distance value is multiplied by the value of fold change for each miRNA. This helps to explore the search space using two different similarity measures based on the distance and ratio of expression values. The power P of fold change varies from 0.1 to 1.5 to find a value which minimizes the average rank of miRNAs for a particular dataset. Here, multiplying two linear functions results in a quadratic function and applying weight as power increases its non linearity beyond the quadratic relation, which helps in identifying non-linear relationship among data points. EDWFC can be described in the following steps:

Step 1: Load the control and resistant group data for a single cancer type. Suppose these are represented by c and r respectively.

Step 2: If there are N number of miRNAs in the dataset, then compute the mean expression value of each miRNA for both the control and resistant groups as follows:

$$M_{cx} = \frac{1}{l}\sum_{i=1}^{l} c_{ix} \tag{2.1}$$

$$M_{rx} = \frac{1}{l}\sum_{i=1}^{l} r_{ix} \tag{2.2}$$

where l is the total number of patients in a group/class. c_{ix} and r_{ix} are the expression value of $i\,th$ patient corresponding to the $x\,th$ miRNA in the control and resistant groups, respectively.

Step 3: Compute the Euclidean distance (D_x) between each mean of the control and resistant group of $x\,th$ miRNA as

$$D_x = \sqrt{(M_{cx} - M_{rx})^2} \tag{2.3}$$

Step 4: Determine fold change of each miRNA as:

$$F_x = \frac{(M_{cx} - M_{rx})}{M_{cx}} \quad if \quad (M_{cx} - M_{rx}) > 0 \tag{2.4}$$

$$or, \quad F_x = \frac{(M_{rx} - M_{cx})}{M_{rx}} \quad if \quad (M_{rx} - M_{cx}) > 0. \tag{2.5}$$

here, F_x is fold change for the $x\,th$ miRNA.

Step 5: Multiply the values of distance and fold change obtained for each miRNA by varying the power P of fold change from 0.1 to 1.5 in steps of 0.1 using m_x as follows:

$$m_x = D_x * (F_x)^P, \quad 0.1 <= P <= 1.5 \tag{2.6}$$

Step 6: Sort the miRNAs in descending order according to the value of m_x. Let R_j represent the rank of the jth miRNA for a particular dataset, considering the set of miRNAs to be R as those reported in the corresponding article.

Step 7: Identify the reported miRNAs ($set \quad R$) in the ranked list obtained in Step 6.

Step 8: Compute the average of the ranks of miRNAs for each value of P (see Step 5) as follows:

$$A = \frac{1}{|R|} \sum_{j \in R} j \tag{2.7}$$

where A is the average of the ranks of all the identified miRNAs in a dataset for a particular value of P and j is the rank of a miRNA. For example, suppose that three miRNAs are identified. For $P = 0.1$, the ranks of those miRNAs are 4, 7, and 10. Then, the average (A) of ranks is 7.

2.3.2 HCEDFCR

The histogram based clustering and the Euclidean distance with fold change based ranking (HCEDFCR) provides groups of miRNAs instead of user specified choice of number of miRNAs. A group of useful miRNAs is required if the user has not determined how many miRNAs will be enough for identifying a drug resistant patient. Further, the ranking of groups and ranking of miRNAs within each group will help to find the important miRNAs. The steps of HCEDFCR are discussed and the pseudocode is provided in Algorithm 2. In HCEDFCR, for a particular dataset, the size of the control group is determined. The square root of the number of rows (miRNAs) is used as the number of bins (B). Here, B is $\sqrt{N}$ such that the number of clusters remains less than equal to $\sqrt{\frac{N}{2}}$ while identifying the valleys in the histogram in a later step. Choosing $\sqrt{\frac{N}{2}}$ number of clusters for a dataset is a widely accepted simple approach [141]. The sum of the expression values of each miRNA in the control data is computed and a histogram is generated using the resultant sum values. The histogram obtained using the sum of expression values helps in identifying the miRNAs with similar magnitude of expression values considering all patients. A valley concept is introduced to define the boundaries for clustering the miRNAs and groups with similar sum of expression values for all patients are identified. A bin is then identified as a valley if its left and right bins are higher (i.e., they have a greater number of miRNAs). This way, all the valleys

Algorithm 1 EDWFC

input : Control, X_x, and drug resistant, Y_x, miRNA expressions.
output: Ranked list of miRNAS

```
Info(Number of miRNAs are a. No. of control and drug resistant patients are b and d, respectively.)
X_x = {x_x1, x_x2, ..., x_xb};
    Y_x = {y_x1, y_x2, ..., y_xd};
    /* x ∈ a                                                              */
Z_k = {z_1, z_2, ..., z_k};
    /* no. of identified miRNAs are k.                                    */
m_cx ← mean(X_x);
  m_rx ← mean(Y_x);
  for x ← 0 to a do
    D_x ← sqrt((m_cx − m_rx)^2);
      if m_cx ≥ m_rx then
        F_x ← (m_cx − m_rx)/m_cx;
    else
        F_x ← (m_rx − m_cx)/m_rx;
    end
end
P = 0.1;
  while P ≤ 1.5 do
    for x ← 0 to a do
        m_x ← D_x * F_x^P;
    end
    rank ← sort(m_x);
      for i ← 0 to k do
        k_Pi ← rank(z_i);
            /* k_P represents the rank of identified miRNAs at value P    */
    end
    avg_P ← (avg(k_P));
       /* avg_P represents the average rank of identified miRNAs at value P */
    P+ = 0.1;
end
min_avg_rank = min(avg_P);
  Return P, rank, min_avg_rank;
```

are located in the histogram. If there are V valleys, then the dataset is divided into $V + 1$ clusters such that the first cluster consists of miRNAs starting from the first bin and ending in the first valley by including those in first valley also. The second cluster consists of miRNAs in the bins after the first valley to the second valley and so forth for the remaining clusters. Finally, the last cluster consists of miRNAs from the bins after the last valley and up to the last bin. For each cluster, the mean of expression values of the control and resistant group are computed and the Euclidean distance between the mean values is determined. Fold change for each miRNA is calculated and multiplied with the corresponding Euclidean distance. The motivations for these steps are similar to those mentioned for EDWFC method. In summary, these similarity measures help in identifying non-linear relationships among data points. The values obtained after the multiplication are sorted in descending order for each cluster to rank the miRNAs within the clusters. The mean of these values is computed for each cluster and sorted in descending order to rank the clusters. HCEDFCR is described as follows:

Step 1: Load the control (c) and resistant (r) data.

Step 2: Find the size of the control dataset. Say, it has N rows (miRNAs) and l columns (patients).

Step 3: Find the square root of N as follows:

$$r = \sqrt{N} \tag{2.8}$$

Step 4: Round the value of the square root calculated in Eq. 8 to be set as B and use it as the number of bins in the histogram.

Step 5: Compute the sum of the expression values for $x\,th$ miRNA in the control group data as:

$$S_x = \sum_{i=1}^{l} c_{ix} \tag{2.9}$$

where S_x is the sum of the expression values of $x\,th$ miRNA in the control group and c_i is the $i\,th$ expression of that miRNA. Here, l is the number of patients in the control group.

Step 6: Plot a histogram using the values of (S_x).

Step 7: Locate the position of valleys in the histogram by identifying bins whose left and right bins are higher.

Step 8: For V number of valleys, divide the control group into $V + 1$ clusters.

Step 9: For each cluster, find the mean of expression values of the control and the resistant group as:

$$M_{cx} = \frac{1}{l} \sum_{i=1}^{l} c_{ix}, \quad and \tag{2.10}$$

$$M_{rx} = \frac{1}{l} \sum_{i=1}^{l} r_{ix} \quad respectively. \tag{2.11}$$

where l is the number of patients in the control or resistant group of a dataset and the variables c_{ix} and r_{ix} are the expression values of $x\,th$ miRNA in the control and resistant groups, respectively.

Step 10: Determine the Euclidean distance between each mean value of the control and resistant group of $x\,th$ miRNA in a cluster as:

$$D_x = \sqrt{(M_{cx} - M_{rx})^2} \tag{2.12}$$

Step 11: For each cluster with w miRNAs, compute fold change of each miRNA as follows:

$$\begin{aligned} F_x &= \frac{(M_{cx} - M_{rx})}{M_{cx}} \quad if \quad (M_{cx} - M_{rx}) > 0 \\ &= \frac{(M_{rx} - M_{cx})}{M_{rx}} \quad if \quad (M_{rx} - M_{cx}) > 0. \end{aligned} \tag{2.13}$$

Step 12: Multiply the value of the Euclidean distance with the value of fold change for $x\,th$ miRNAs in a cluster as:

$$m_x = D_x * (F_x) \tag{2.14}$$

Step 13: Sort the values of m_x in a descending order for each cluster to rank the miRNAs within it.

Step 14: To rank the clusters, compute the mean of the values obtained in Step 12 of $k\,th$ cluster as:

$$M_k = \frac{1}{w} \sum_{m_x \in k} [m_x] \tag{2.15}$$

where w is the number of miRNAs in the k^{th} cluster.

Step 15: Sort M_k in descending order to rank the clusters.

The source code, for EDWFC and HCEDFCR, and the steps to run the code are provided on Github page `https://github.com/joginder12/EDWFC-HCEDFCR`.

2.4 Experimental Evaluations

The methods, EDWFC and HCEDFCR, are implemented in the Python programming language using Intel(R)Xeon(R) CPU E3-1270 V2 @ 3.50 GHz processor and 32 GB RAM. The EDWFC provides a ranked list of miRNAs, whereas the HCEDFCR delivers clusters of miRNAs with cluster ranks. EDWFC and HCEDFCR are evaluated in terms of their capabilities to identify the top miRNAs associated with drug resistance as well as in terms of comparative results w.r.t. related methods. The miRNAs identified in the existing biochemical studies as drug resistant ones are used to validate the top miRNAs identified by any method.

2.4.1 Evaluation of miRNAs selected by EDWFC

Table 2.2 shows the best ranks of the miRNAs achieved by EDWFC for various datasets. The power P of fold change corresponding to those ranks and the average of ranks for

Algorithm 2 HCEDFCR

input : Control, X_x, and drug resistant, Y_x, miRNA expressions.
output: A group of miRNAs associated with drug resistance.

Info(*Number of miRNAs are a. No. of control and drug resistant patients are b and d, respectively.*)
$X_x = \{x_{x1}, x_{x2}, \ldots, x_{xb}\}$;
$Y_x = \{y_{x1}, y_{x2}, \ldots, y_{xd}\}$;
/* $x \in a$ */
$m_{cx} \leftarrow mean(X_x)$;
$m_{rx} \leftarrow mean(Y_x)$;
$B \leftarrow \sqrt{a}$;
/* Rounding off of B provides no. of valleys in histogram. */
$S_x \leftarrow sum(X_x)$;
Plot(*Plot histogram using S_x.*)
Identify valleys.;
if *valleys* $== V$ **then**
  $n_{cluster} \leftarrow V + 1$;
    /* $n_{cluster}$ represents the no. of clusters. */
  $cluster_{label} \leftarrow \{0, 1, \ldots, v\}$;
end
for $x \leftarrow 0$ **to** a **do**
  $D_x \leftarrow \sqrt{(m_{cx} - m_{rx})^2}$;
  if $m_{cx} \geq m_{rx}$ **then**
    $F_x \leftarrow (m_{cx} - m_{rx})/m_{cx}$;
  else
    $F_x \leftarrow (m_{rx} - m_{cx})/m_{rx}$;
  end
  $m_x \leftarrow D_x * F_x$;
end
for $v \leftarrow 0$ **to** V **do**
  for $x \leftarrow 0$ **to** a **do**
    if $Cluster_{label}(S_x) == v$ **then**
      $f_{vx} \leftarrow m_x$;
    end
  end
end
for $v \leftarrow 0$ **to** V **do**
  $M_k \leftarrow sum(f_v)/c_v$;
    /* c_v is the no. of elements in cluster v */
end
Sort(M_k);
$top_{cluster_{rank}} \leftarrow min(M_K)$;

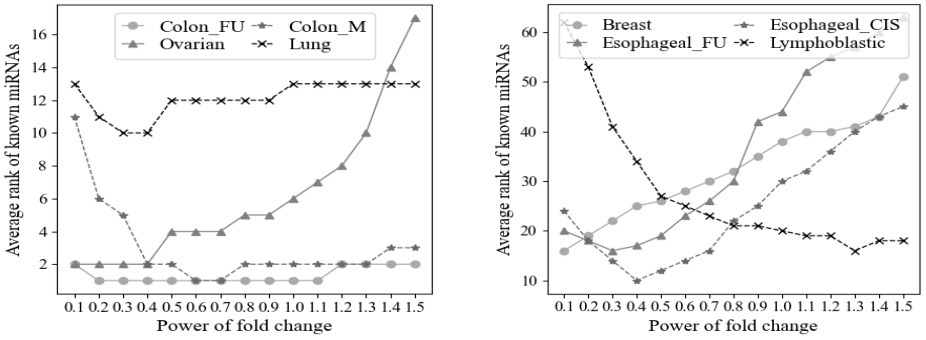
a particular dataset are also reported in the table. Only in cases of 2 out of 8 datasets, the average rank of miRNAs goes beyond 10 (the average rank of known miRNAs is 16 for both breast can lymphoblastic datasets). For all the datasets, the reported miRNAs lie within the top 25 of the ranked list. Interestingly, the miRNAs in Colon_FU and Colon_M are ranked at top of the list and the miRNA in ovarian cancer is ranked second. The P value and average rank for breast, Colon_FU, Colon_M, Esophageal_CIS, Esophageal_FU, lung, ovarian and lymphoblastic leukemia datasets are 0.1 and 16, 0.2 and 1, 0.6 and 1, 0.4 and 10, 0.4 and 5, 0.3 and 10, 0.1 and 2, and 1.3 and 16, respectively.

Figs. 2.3(a) and 2.3(b) show the variation in the average rank of miRNAs for different powers of fold change. Here, the x-axis shows the power P of fold change while the y-axis

Table 2.2: Evaluations of selected miRNAs by EDWFC.

Cancer Type	miRNA names	Best ranks	value of P and average rank
Breast	miR-21	17	0.1, 16
	miR-103	19	
	miR-221	05	
	miR-222	25	
Colon_FU	miR-19b	1	0.2, 1
Colon_M	miR-224	1	0.6, 1
Esophageal_CIS	miR-125a-5p	5	0.4, 10
	miR-455-3p	9	
	miR-638	17	
Esophageal_FU	miR-125a-5p	5	0.4, 5
	miR-935	6	
Lung	miR-134	11	0.3, 10
	miR-487b	7	
	miR-655	13	
Lymphoblastic	let-7b	15	1.3, 16
	miR-93	23	
	miR-126*	17	
	miR-191	3	
	miR-213	21	
	miR-223	2	
Ovarian	miR-634	2	0.1, 2

shows average ranks of miRNAs associated with drug resistance. The average ranks in Fig. 2.3(a) correspond to Colon_FU, Colon_M, ovarian, and lung data, while Fig. 2.3(b) provides the same for breast, Esophageal_FU, Esophageal_CIS, and lymphoblastic.



(a) (b)

Figure 2.3: Variation of 'average rank of reported miRNAs' with power (P) of fold change using EDWFC. (a) Curves for Colon_FU, Colon_M, Ovarian, and Lung datasets. (b) Curves for Breast, Esophageal_CIS, Esophageal_FU, and Lymphoblastic Leukemia datasets.

2.4.2 Performance Evaluation of EDWFC

The efficiency of the developed EDWFC is evaluated by comparing it with some existing feature selection methods like SPEM [37], 1D-CNN [131], CSPS [138], MRMR [33], SVM-RFE [38], SVMRFE with MRMR (SVMRFE_M) [98], FSNLDA [99], and Lasso [139]. These methods are used for miRNA or gene expression analysis in cancer. All the above

mentioned approaches are either classical and highly cited methods or recently developed techniques for miRNA or gene selection using their expression values. First, the miRNAs are ranked using any particular method and the top 1% miRNAs are selected from the ranked list. The classification performance of the selected (top 1%) miRNAs in terms of specificity, sensitivity, accuracy, F score, and MCC using the support vector machine (SVM), random forest (RF), and Naive Bayes (NB) classifiers is shown in Table 2.3. Recently, boosted trees gained popularity and were also used for binary classification in several studies. However, in this investigation, the number of samples is low, and the boosted trees may suffer from overfitting [141] issues. Hence, the above mentioned classifiers are used for classifying control and drug resistant patients using selected miRNAs. Here, leave one out cross-validation technique is used for the performance measurement. The results in terms of minimum and maximum values of F scores are 0.72 and 1.0, 0.73 and 1.0, and 0.67 and 1.0 for SVM, Naive Bayes (NB), and random forest (RF) classifiers, respectively, for different datasets. Similar results are also obtained in terms of specificity, sensitivity, accuracy, and MCC and shown in Table 2.3. To show the efficiency of EDWFC, the classification performance using the top 1% miRNAs obtained from each method are then compared using the support vector machine (SVM) and random forest (RF) classifier in terms of their classification accuracy in discriminating the control (cancer patients without drug) and drug resistant cancer patients.

Table 2.3: Classification results for top 1% miRNAs selected by EDWFC using support vector machine (SVM), random forest (RF), and Naive Bayes (NB) classifiers of various cancer datasets.

Classifiers	Datasets	Sensitivity	Specificity	Accuracy	F-score	MCC
SVM	Breast	1	1	100	1	1
	Colon_FU	0.81	0.71	76.19	0.77	0.52
	Colon_M	0.63	0.81	67.86	0.74	0.40
	Esophageal_FU	0.91	0.88	89.58	0.89	0.79
	Esophageal_CIS	0.91	0.88	89.58	0.89	0.79
	Lung	0.75	0.75	75.00	0.75	0.50
	Lymphoblastic	0.75	0.71	72.99	0.72	0.46
RF	Breast	1	1	100	1	1
	Colon_FU	0.74	0.79	71.88	0.74	0.44
	Colon_M	0.71	0.63	66.67	0.68	0.35
	Esophageal_FU	0.90	0.89	89.82	0.90	0.80
	Esophageal_CIS	0.91	0.89	89.84	0.90	0.80
	Lung	0.75	0.75	75.00	0.75	0.50
	Lymphoblastic	0.65	0.70	67.68	0.67	0.36
NB	Breast	1	1	100	1	1
	Colon_FU	0.76	0.78	72.76	0.75	0.48
	Colon_M	0.76	0.67	68.35	0.73	0.43
	Esophageal_FU	1.0	0.83	91.67	0.92	0.85
	Esophageal_CIS	1.0	0.83	91.67	0.92	0.85
	Lung	0.75	0.75	75.00	0.75	0.50
	Lymphoblastic	0.72	0.74	74.14	0.76	0.48

The results using SVM as a classifier are presented in Table 2.4, where the best results are marked in bold. From the table, it is evident that our method provides comparable or better results than the related techniques used for comparison in all the cases except

Table 2.4: Comparing EDWFC with different methods using SVM classifier. The best results are marked in bold.

Datasets	Methods	Sensitivity	Specificity	Accuracy	F-score	MCC
Breast	EDWFC	**1**	**1**	**100**	**1**	**1**
	SPEM	1	1	100	1	1
	CNN	1	1	100	1	1
	MRMR	1	1	100	1	1
	SVMRFE	0.97	0.92	95.23	0.95	0.90
	SVMRFE_M	0.97	0.92	95.23	0.95	0.90
	FSNLDA	1	0.85	92.85	0.92	0.87
	CSPS	1	1	100	1	1
	Lasso	0.96	0.93	95.50	0.94	0.90
Colon_FU	EDWFC	**0.81**	0.71	**76.19**	**0.77**	**0.52**
	SPEM	0.79	0.56	62.50	0.66	0.38
	CNN	0.72	**0.80**	75.0	0.76	0.51
	MRMR	0.24	57	40.48	0.34	-0.20
	SVMRFE	0.52	0.22	52.38	0.52	0.04
	SVMRFE_M	0.53	0.23	52.35	0.53	0.05
	FSNLDA	0.57	0.57	57.14	0.57	0.11
	CSPS	0.62	0.42	52	0.50	0.04
	Lasso	0.53	0.25	53.36	0.54	0.05
Colon_M	EDWFC	0.63	**0.81**	67.86	**0.74**	0.40
	SPEM	**0.84**	0.58	63.80	0.68	0.41
	CNN	0.60	0.73	70.83	0.67	0.37
	MRMR	0.71	0.71	**71.42**	0.71	**0.43**
	SVMRFE	0.53	0.42	48.21	0.47	-0.03
	SVMRFE_M	0.54	0.43	49.34	0.48	-0.01
	FSNLDA	0.35	0.42	39.38	0.21	-0.03
	CSPS	0.54	0.50	51.78	0.52	0.03
	Lasso	0.56	0.45	53.75	0.51	0.07
Esophageal_FU	EDWFC	**0.91**	0.88	**89.58**	**0.89**	**0.79**
	SPEM	0.83	**0.92**	87.50	0.87	0.73
	CNN	0.89	0.73	83.44	0.60	0.76
	MRMR	0.88	0.90	89.28	0.89	0.79
	SVMRFE	0.37	0.52	44.79	0.43	-0.10
	SVMRFE_M	0.38	0.53	44.78	0.44	-0.10
	FSNLDA	0.79	0.81	80.20	0.60	0.66
	CSPS	0.54	0.58	56.25	0.56	0.12
	Lasso	0.78	0.82	81.35	0.79	0.65
Esophageal_CIS	EDWFC	**0.91**	0.88	**89.58**	**0.89**	**0.79**
	SPEM	0.83	**0.92**	87.50	0.87	0.73
	CNN	0.91	0.69	80.64	0.58	0.71
	MRMR	0.79	0.87	83.33	0.83	0.67
	SVMRFE	0.33	0.52	42.86	0.40	-0.14
	SVMRFE_M	0.33	0.52	42.86	0.40	-0.14
	FSNLDA	0.79	0.81	80.20	0.80	0.60
	CSPS	0.54	0.58	56.25	0.56	0.12
	Lasso	0.78	0.82	81.35	0.79	0.65
Lung	EDWFC	**0.75**	**0.75**	**75.00**	**0.75**	**0.50**
	SPEM	0.61	0.72	66.57	0.59	0.32
	CNN	0.66	0.71	68.75	0.70	0.40
	MRMR	0.42	0.33	37.50	0.37	-0.25
	SVMRFE	0.25	0.58	41.67	0.35	-0.18
	SVMRFE_M	0.24	0.56	41.65	0.34	-0.17
	FSNLDA	0.68	0.50	59.37	0.66	0.19
	CSPS	0.38	0.43	40.63	0.40	-0.18
	Lasso	0.65	0.56	62.50	0.67	0.21
Lymphoblastic	EDWFC	0.75	0.71	72.99	0.72	0.46
	SPEM	0.74	0.69	70.67	0.66	0.34
	CNN	**0.83**	**0.92**	**91.37**	**0.91**	**0.8**
	MRMR	0.63	0.63	63.21	0.63	0.26
	SVMRFE	0.55	0.65	60.34	0.60	0.21
	SVMRFE_M	0.54	0.63	59.99	0.65	0.23
	FSNLDA	0.59	0.61	60.34	0.60	0.21
	CSPS	0.56	0.63	59.48	0.59	0.19
	Lasso	0.62	0.58	59.35	0.59	0.19

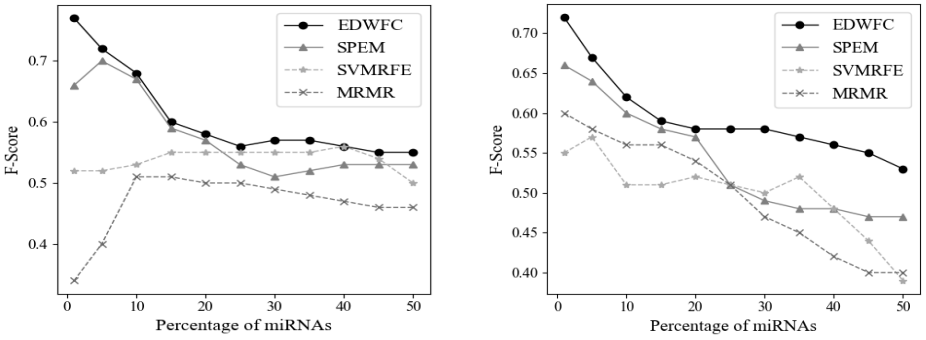
Table 2.5: Comparing EDWFC with other methods using random forest classifier. The best results are marked in bold.

Datasets	Methods	Sensitivity	Specificity	Accuracy	F-score	MCC
Breast	EDWFC	**1**	**1**	**100**	**1**	**1**
	SPEM	0.98	0.95	96.61	0.98	0.95
	CNN	1	1	100	1	1
	MRMR	0.93	0.97	95.00	0.95	0.90
Colon_FU	EDWFC	0.74	0.79	71.88	0.74	0.44
	SPEM	**0.79**	0.56	62.50	0.66	0.38
	CNN	**0.80**	0.66	70.32	0.74	**0.47**
	MRMR	0.53	0.35	43.75	0.48	-0.13
Colon_M	EDWFC	0.71	0.63	66.67	0.68	0.35
	SPEM	0.66	0.53	60	0.62	0.18
	CNN	**0.73**	0.66	68.75	0.65	0.38
	MRMR	0.70	**0.74**	**72.22**	**0.72**	**0.45**
Esophageal_FU	EDWFC	**0.90**	**0.89**	**89.82**	**0.90**	**0.80**
	SPEM	0.90	0.88	88.88	0.89	0.77
	CNN	0.88	0.72	80.56	0.82	0.63
	MRMR	0.83	0.80	81.67	0.82	0.64
Esophageal_CIS	EDWFC	**0.91**	**0.89**	**89.84**	**0.90**	**0.80**
	SPEM	0.90	0.88	88.89	0.89	0.77
	CNN	0.91	0.74	83.33	0.84	0.66
	MRMR	0.86	0.82	85	0.85	0.70
Lung	EDWFC	**0.75**	**0.75**	**75.00**	**0.75**	**0.50**
	SPEM	0.66	0.68	67.50	0.68	0.37
	CNN	0.62	0.73	65.63	0.69	0.33
	MRMR	0.53	0.49	50.71	0.52	0.02
Lymphoblastic	EDWFC	0.65	0.70	67.68	0.67	0.36
	SPEM	0.67	0.64	66.20	0.67	0.32
	CNN	**0.81**	**0.87**	**83.62**	**0.84**	**0.65**
	MRMR	0.49	0.51	50.68	0.50	0.02

for 1D-CNN using lymphoblastic leukemia data (0.83, 0.92, 91.37, 0.91, and 0.80) and using Colon_FU data in terms of specificity (0.80), for SPEM using Colon_M data in terms of sensitivity (0.84) and two types of esophageal data in terms of specificity (0.92 and 0.92), and for MRMR using Colon_M data in terms of accuracy (71.42) and MCC (0.43). In other words, EDWFC provides the best results in 302 out of 315 cases (7 datasets $\times$ 9 methods $\times$ 5 measures $\times$ 1 classifier). In the case of the lymphoblastic dataset, 1D-CNN provides the best results, but not in other datasets where the sample space is low and an overfitting problem arises.

In Table 2.5, EDWFC is compared with SPEM, 1D CNN, and MRMR using random forest as a classifier. Here, SPEM, 1D CNN and MRMR are chosen as they are within the top three methods when SVM is used as a classifier. From the table, it is clear that EDWFC performs better than the compared methods except for 1D CNN using lymphoblastic data for all the measures and using Colon_FU data in terms of sensitivity and MCC, for SPEM using Colon_M and lymph data in terms of sensitivity, and for MRMR using Colon_M data in terms of specificity, accuracy, f-score and MCC. Here, EDWFC provides the best results in 127 out of 140 cases (7 datasets $\times$ 4 methods $\times$ 5 measures $\times$ 1 classifier).

The EDWFC is also compared with the related methods using various percentages of



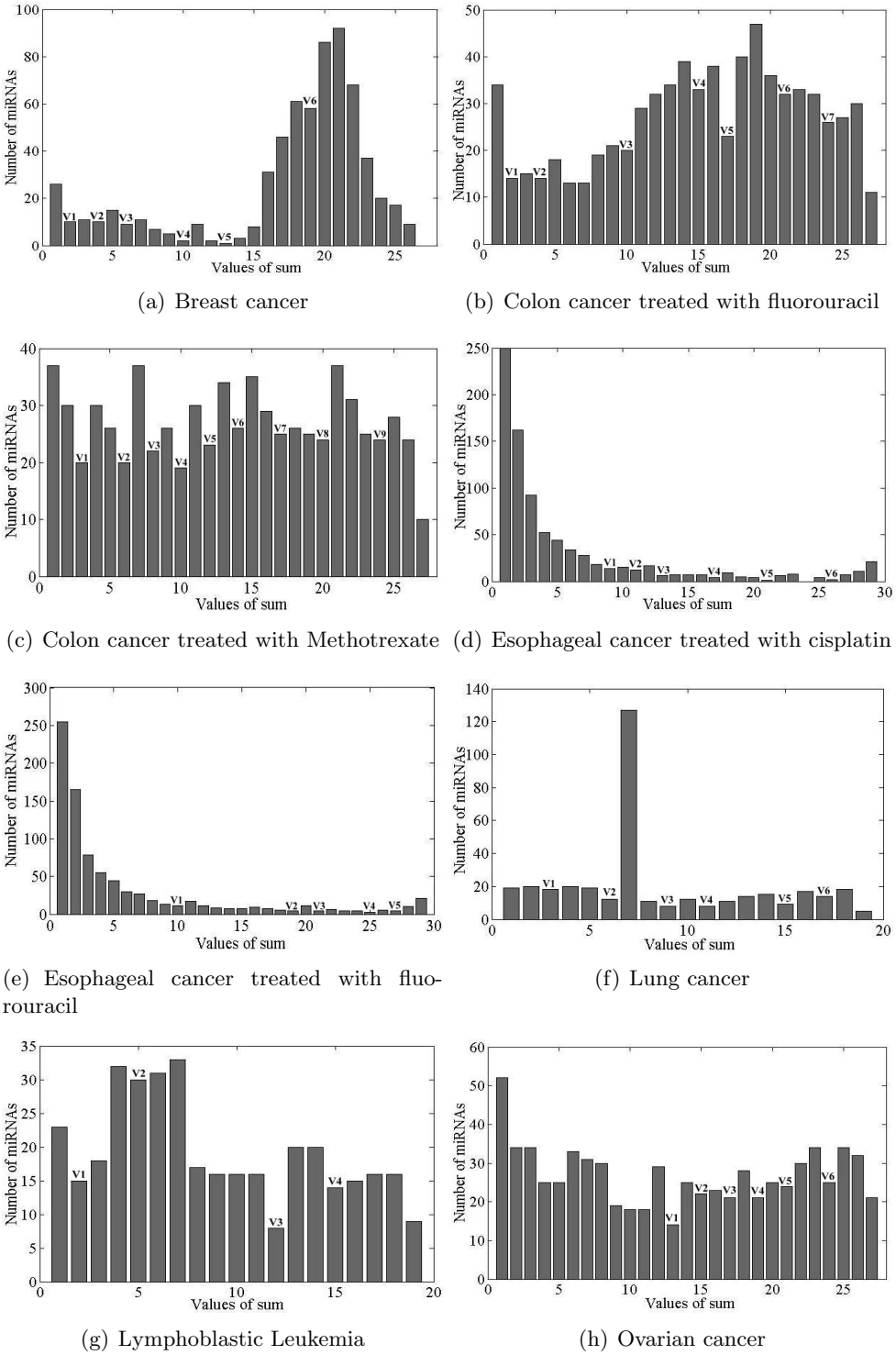
(a) Colon Cancer treated with fluorouracil (b) Lymphoblastic Leukemia

Figure 2.4: Variation of F-score with different percentages of miRNAs for various methods. SVM is used as a classifier.

selected miRNAs in terms of F-score. Figs. 2.4(a) and 2.4(b) show the related curves for colon cancer treated with fluorouracil drug and lymphoblastic leukemia datasets using EDWFC. SVM is used as a classifier. The curves for EDWFC lie at the top in both figures. Note that SPEM achieved comparable F-score values with EDWFC using 15% and 20% percent of miRNAs for colon cancer treated with fluorouracil and lymphoblastic leukemia data. Further, the F-score decreases at higher percentages of miRNAs.

2.4.3 Evaluation of miRNAs selected by HCEDFCR

In this section, the results obtained by HCEDFCR for the different cancer datasets are discussed. The nature of the bins and valleys corresponding to each dataset is shown in Fig. 2.5(a)-2.5(h). Here, the bins identifying the valleys are marked as $V_1, V_2, \ldots, V_n$. The numbers of valleys are 6, 7, 9, 6, 5, 6, 4, and 6 for breast, colon_FU, colon_M, esophageal_CIS, esophageal_FU, lung, lymphoblastic and ovarian datasets, respectively. These valleys result in 7, 8, 10, 7, 6, 7, 5, and 7 clusters for the same datasets. The minimum and maximum number of clusters are 10 and 5 for colon_M and lymphoblastic datasets, respectively, (Figs. 2.5(c) and 2.5(g)). The results of HCEDFCR are provided in columns 1 to 4 of Table 2.6. The remaining columns (5, 6, and 7) will be required in the next section (Section 2.4.4) for comparison with related methods. For each dataset, the number of clusters, names of the miRNAs, the rank of the clusters in which the miRNAs are found along with the ranks of those miRNAs in the cluster are reported. Interestingly, miR-19b in Colon_FU, miR-224 in Colon_M, miR-634 in ovarian dataset are ranked at the first position in the top cluster. For breast cancer, miRNAs are found within the top three ranking positions in the top four clusters. For example, miR-21 and miR-221 are ranked second and third in the first and second clusters, respectively, whereas miR-222 and miR-103 are ranked second and third in the fourth cluster. In



(a) Breast cancer

(b) Colon cancer treated with fluorouracil

(c) Colon cancer treated with Methotrexate

(d) Esophageal cancer treated with cisplatin

(e) Esophageal cancer treated with fluorouracil

(f) Lung cancer

(g) Lymphoblastic Leukemia

(h) Ovarian cancer

Figure 2.5: Identification of valleys in the histogram using HCEDFCR for various datasets.

lymphoblastic data, all miRNAs are found in the top 2 clusters except miR-93 which is ranked first in the fourth cluster.

Table 2.6: Evaluations of selected miRNAs by HCEDFCR, spectral clustering, k-means, and SOM.

Cancer Type	Total no. of clusters	miRNA names	cluster rank with miRNA rank *HCED FCR*	cluster rank with miRNA rank *Spectral Clustering*	cluster rank with miRNA rank $K-means$	cluster rank with miRNA rank *SOM*
Breast	07	miR-21	01-02	01-01	03-82	06-61
		miR-103	04-03	04-05	03-55	06-37
		miR-221	02-03	02-03	03-26	01-23
		miR-222	04-02	04-11	02-28	05-37
Colon_FU	08	miR-19b	01-01	04-379	01-166	02-154
Colon_M	10	miR-224	01-01	03-234	01-181	01-174
Esophageal_CIS	07	miR-125a-5p	03-03	01-194	01-135	01-142
		miR-455-3p	03-02	01-107	01-26	01-26
		miR-638	03-01	04-201	01-227	01-72
Esophageal_FU	06	miR-125a-5p	03-03	01-194	01-135	05-45
		miR-935	02-01	01-279	01-45	01-61
Lung	07	miR-134	01-02	03-14	01-22	03-29
		miR-487b	01-03	03-36	02-107	05-31
		miR-655	02-02	06-22	01-35	03-15
Lymphoblastic	05	let-7b	01-01	02-13	05-23	01-29
		miR-93	04-01	02-43	03-25	04-17
		miR-126*	01-06	02-46	02-52	01-12
		miR-191	01-03	04-17	02-42	01-47
		miR-213	01-08	04-49	05-33	03-13
		miR-223	02-01	04-60	04-21	01-65
Ovarian	07	miR-634	01-01	02-1	01-55	01-52

2.4.4 Performance Evaluation of HCEDFCR

The HCEDFCR is compared with three popular clustering algorithms: spectral clustering [29], Kmeans [158], and self-organizing maps (SOM) [27]. The radial basis function (RBF) is used for similarity matrix computation in spectral clustering. The k-means and SOM use the Euclidean distance as a metric for similarity calculation. For comparison purposes, thehistogram-basedd clustering is replaced by each of these clustering methods in HCEDFCR, while the steps such as the ranking of clusters and the ranking of miRNAs inside the clusters are kept the same.

The comparison results of HCEDFCR with spectral clustering, Kmeans, and SOM are provided in Table 2.6. The miRNAs obtained using spectral clustering are within the top 10 ranks for top 4 clusters only for 2 datasets, breast and ovarian. Hence, the results of spectral clustering are close to HCEDFCR for only two datasets and inferior for others. For example, using ovarian dataset, while miR-634 is ranked second in the first cluster using spectral clustering, it is ranked first using HCEDFCR. The results of k-means and SOM show that the miRNAs are poorly ranked in all the datasets. In some cases, SOM performs better than k-means and spectral clustering but is inferior to HCEDFCR. For example, using lymphoblastic data, miR-126* and miR-213 are ranked 12th in cluster 1 and 13th in cluster 3, respectively using SOM, whereas they are ranked 6th and 8th in the first cluster using HCEDFCR.

The results of related methods are not close to the desired level and are inferior to HCEDFCR. A possible reason may be that for the drug resistant problem, the overall magnitude of expression values across all control patients for a miRNA is more important than the expression value of each patient. This is accomplished in HCEDFCR by considering the sum of expression values (Step 5, Eq. 2.9) for a miRNA in the control group.

2.5 Complexity of EDWFC and HCEDFCR

In this section, the time complexity of EDWFC and HCEDFCR is discussed where the expression data consists of l number of patients in a group/class and N number of miRNAs in a cancer type.

2.5.1 Complexity of EDWFC

Suppose l number of patients are present in both the control and drug resistant classes. Each patient will have one expression value of a miRNA. Therefore, the time to compute the mean of expression values for each miRNA for both classes will be $O(l + l)$, which can be represented as $O(l)$. The time to compute the mean of expression values for N miRNAs in a dataset will be $O(l \times N)$. The Euclidean distance and weighted fold change (varying the power p from 0.1 to 1.5) can be calculated in $O(N)$ and $O(p \times N)$. For each value of p, the miRNAs in the list are sorted with complexity $O(p \times N \log N)$ time. The total computational complexity for EDWFC can be computed as $O(l \times N + N + p \times N + p \times N \log N)$.

2.5.2 Complexity of HCEDFCR

In the previous section, l and N are the number of samples in a group/class and the number of miRNAs, respectively. The sum of expression values of a miRNA is calculated for a class, and the required time is $O(l)$. It is performed for all the miRNAs, and the time is $O(l \times N)$. The summed values are distributed in B number of bins into a histogram with computational complexity time $O(N + B)$. The complexity to detect valleys in a histogram is $O(B)$.

The miRNAs will be assigned to each cluster with complexity $O(N)$. The mean of expression values can be computed in $O(l \times N)$. The ED and FC can be calculated in $O(N)$. The average EDFC score for each cluster is computed with complexity $O(N)$. The clusters are sorted based on their average EDFC score, and the complexity for this operation is $O(C \log C)$. The complexity of sorting miRNAs inside clusters can be at

most $\sum_{i=1}^{C} O(m_i \log m_i) \leq O(m \log m)$. Therefore, the overall computational complexity of HCEDFCR is $O(l \times N + N + B + C \log C + m \log m)$.

2.6 Discussion and Conclusion

In this chapter, two methods (EDWFC and HCEDFCR) are presented to identify a subset of miRNAs associated with drug resistance in patients with cancer. In the EDWFC, the Euclidean distance and fold change between the averages of control and resistant expressions are multiplied by varying the power of the fold change value to minimize the average rank of biologically known miRNAs. A portion of miRNAs is selected from the top of the ranked list. In the HCEDFCR, a histogram-based clustering method is employed to categorize miRNAs into distinct clusters. The product of the Euclidean distance and fold change value between the control and resistant miRNA expressions is used to rank the clusters and miRNAs inside the clusters. MiRNAs in the top-ranked cluster are considered as important.

The results of EDWFC show that miRNAs associated with drug resistance are present within the top 20 positions in the ranked list for all the datasets except for breast cancer, where one of the identified miRNAs is ranked 25th. Moreover, the important miRNAs, as reported in the corresponding biochemical studies, are found within the top five positions in the ranked list for all the datasets. The classification performance of EDWFC, using top 1% miRNAs from the ranked list, is compared with related methods and EDWFC provided the best results in 302 out of 315 cases and 127 out of 140 cases, using SVM and RF classifiers, respectively. The variations of F-score for different percentages of miRNAs using various methods for two datasets (colon cancer treated with fluorouracil drug and lymphoblastic leukemia datasets) show that the performance of all the methods decreases as we increase the percentage of miRNAs.

From the results of HCEDFCR, it is observed that the top three miRNAs of the top four clusters contain the miRNAs associated with drug resistance for all the datasets, except for the lymphoblastic leukemia dataset where the miRNAs are found in the sixth and eighth position of cluster 1. Comparison of HCEDFCR with other clustering methods shows that HCEDFCR performs better in identifying important drug resistant miRNAs. In some cases such as breast, ovarian, and lymphoblastic datasets, spectral clustering and SOM perform well and their results are close but inferior to HCEDFCR. In general, the results indicate that HCEDFCR can serve the purpose of detecting the miRNAs associated with drug resistance in patients with cancer.

In summary, if the user wants to get a specific number of miRNAs for a particular cancer and related drug, then the EDWFC is recommended to rank individual miRNAs based on its capability to differentiate between two classes (control and drug resistance). The user then selects the required number of miRNAs from the top of the ranked list.

On the other hand, if the user wants a group of miRNAs without any user defined number of miRNAs, then the HCEDFCR is recommended. The HCEDFCR provides the ranked clusters. The miRNAs in the top ranked clusters are considered as groups of important miRNAs. In future, the EDWFC and HCEDFCR can be used on similar miRNA datasets where the miRNAs from the top of the ranked list and the top five miRNAs from the top clusters, respectively, can be biochemically tested to identify the miRNAs associated with cancer drug resistance.

Chapter 3

Integrating Fuzzy Rough Set-Based Entropies for Identifying Drug Resistant miRNAs and Classifying Cancer Patients

3.1 Introduction

In Chapter 2, all miRNAs, within a dataset, are ranked using the biological knowledge of known drug resistant miRNAs and the performance of the top ranked miRNAs is evaluated in terms of classifying control and drug resistant patients. In this chapter, the miRNAs are ranked using information about patient labels and evaluated in a way similar to that in Chapter 2. A framework, named "a weighted framework for integrating fuzzy rough set-based relevance and redundancy entropies (WFIFRRRE)", for identifying a set of miRNAs that efficiently classifies the control and drug resistant patients, is developed. The weight is introduced for integrating two z score based relevance and redundancy entropies. The entropies are computed using fuzzy membership function and rough approximation (lower and boundary region), two major components of soft computing. While fuzzy set handles the uncertainty emerging from the overlapping expressions of miRNAs (two miRNAs expressions can overlap, and expressions of any miRNA in two samples/patients can also overlap), rough set takes care of the exactness in the class shape of control and drug resistant groups. WFIFRRRE is comprised of four important steps: (a) compute the fuzzy relevance and redundancy membership of each expression value using the z score based fuzzy membership function and determine their upper and lower approximations using indiscernibility relations of rough set, (b)

find the relevance of each miRNA to a particular class (control or drug resistant) by calculating the relevance entropy measure, (c) compute average redundancy entropy of each miRNA w.r.t other miRNAs in the dataset, (d) integrate the relevance entropy and average redundancy entropy through weights in a supervised manner to maximize the accuracy of classifying patients using top-ranked miRNAs. The set of miRNAs responsible for achieving maximum accuracy for patient classification is selected from the ranked list. The novelty of this work lies in i) formulating a new z score based fuzzy membership function to compute the membership of each expression value and then determine the relevance and redundancy entropy measures for each miRNA, and ii) developing a framework for integrating the relevance and redundancy entropies for miRNA ranking and selection.

The classification performance of selected miRNAs by WFIFRRRE and the related methods is evaluated using Naive Bayes (NB), random forest (RF), and SVM classifiers. For cross validation, leave one out technique is used. The top 1% of miRNAs is chosen for all the compared methods to classify the control and drug resistant patients. The methods are evaluated in terms of sensitivity, specificity, accuracy, F score, and MCC. WFIFRRRE outperforms the relevant approaches used for comparison in most of the cases. The miRNAs selected by WFIFRRRE are validated using different existing biochemical/biological studies. Most of the selected miRNAs by WFIFRRRE are mentioned in those investigations.

The chapter is organized into six more sections. An outline of various datasets is provided in 3.2. The methodology is described in 3.4. Expermental results are reported in Section 3.5. The biological relevance of selected miRNAs is discussed in Section 3.6. The study is concluded in Section 3.7.

3.2 Datasets

Eight miRNA expression cancer datasets, viz., esophageal, ovarian, colon M, colon FU, breast, lung, lymphoblastic, and squamous, are downloaded from the publicly available data archive Gene Expression Omnibus (GEO) [94]. The datasets used in the studies are deposited by the researchers to GEO by taking care of the pertinent ethical concerns. The number of miRNAs and patients in different datasets are provided in Table 3.1 along with their source articles. The datasets, except for esophageal and squamous, in this chapter are the same as Chapter 2. Here, while the drug resistant patients of the esophageal and squamous data are the same as those in the Esophageal_FU and Esophageal_M datasets in Chapter 2, the control patients of the Esophageal_FU and Esophageal_M datasets are combined and used in both the esophageal and squamous datasets. This is performed to increase the sample size and is inspired by the study in [37].

The number of control and drug resistant patients is equal for each cancer type except esophageal and squamous cell, where the number of control patients is more than the number of drug resistant patients. The total number of samples is 6 in colon cancer treated with the methotrexate drug (Colon M) and 58 in lymphoblastic leukemia which are minimum and maximum among all the datasets, respectively. Note that, some miRNAs are repeated with different expression values in some of the used datasets, as expressions of these miRNAs are generated using different expression-detecting techniques. The repeated miRNAs are also counted in the total number of miRNAs.

Table 3.1: An outline of the miRNA expression datasets

Cancer Type	No. of miRNAs	No. of Control Patients	No. of Drug resistant Patient
Esophageal [76]	847	12	6
Ovarian [74]	727	7	7
Colon FU [73]	723	4	4
Colon M [72]	723	3	3
Breast [71]	654	6	6
Lung [156]	377	4	4
Lymphoblastic [77]	365	29	29
Squamous [76]	847	12	6

3.3 Preliminary Concepts

In this section, first, some preliminary definitions like crisp set, fuzzy set, and rough set are provided, and then the advanced concepts like fuzzy rough sets and fuzzy rough approximation regions are defined.

Let, U be the universal non-empty finite set and A be the set of attributes then $R =< U, A >$ will be an information system [159]. In our problem, the information system can be represented with a row-column table where the columns will represent patients/samples and the rows will denote miRNAs as objects. Patients are labeled as control or drug resistant. Let us now define the crisp set, fuzzy set, rough set, and fuzzy rough set.

Crisp set: Let, $\chi \subseteq U$ then χ is a crisp set if $\chi(g) \in \{0, 1\}$ where $g \in \chi$ [160].

Fuzzy set: Let, $\chi' \subseteq U$ then χ' is a fuzzy set if $\chi'(g) \in [0, 1]$ where $g \in \chi'$ [160].

Here, $\chi(g)$ and $\chi'(g)$ represent the membership of g in χ and χ', respectively.

Rough Set: Let, β creates a group of similar objects called granules where β is an indiscernibility relation [161]. The indiscernibility among objects results in rough definition of χ. χ can be approximated using indiscernibility relation β by defining

$\beta - lower$ and $\beta - upper$ regions of set χ as [161]:

$$\underline{\beta}\chi = \{g \in \chi' | [g]_\beta \subseteq \chi'\},\ and \tag{3.1}$$

$$\overline{\beta}\chi = \{g \in \chi' | [g]_\beta \cap \chi' \neq \phi\}. \tag{3.2}$$

Here, $\underline{\beta}\chi$ and $\overline{\beta}\chi$ are the lower and upper approximation regions of set χ, respectively. The tuple $< \underline{\beta}\chi, \overline{\beta}\chi >$ is rough set.

Fuzzy Rough Set: From the above definition, if β and χ are a fuzzy indiscernibility relation and a crisp set, respectively, then the tuple $< \underline{\beta}\chi, \overline{\beta}\chi >$ will result in a fuzzy rough set of χ. $\underline{\beta}\chi$ and $\overline{\beta}\chi$ are derived as:

$$\underline{\beta}\chi = \{(g, \underline{\mu}(g)) | g \in U\},\ and \tag{3.3}$$

$$\overline{\beta}\chi = \{(g, \overline{\mu}(g)) | g \in U\}. \tag{3.4}$$

Here, in Eq. (3.3, 3.4), g is an element in U and $\underline{\mu}(g)$ and $\overline{\mu}(g)$ represents the membership of g in $\underline{\beta}\chi$ and $\overline{\beta}\chi$, respectively.

3.4 Weighted Framework for Integrating Fuzzy Rough Set-based Relevance and Redundancy Entropies

The existing fuzzy membership functions do not consider the distance of each expression value from the mean of both classes in terms of standard deviations. Hence, we develop a membership function which not only addresses the aforementioned issue but also considers the belongingness of an expression value in both the classes by computing the ratio of its distance from the means of both the classes. The membership of expression value will be higher in the class in which z score of the expression value is minimum. Further, the existing research works for finding rank of miRNAs, to properly discriminate patients deal with sequential computation of first relevant miRNAs and then redundant miRNAs. These approaches cannot handle both relevance and redundancy of a miRNA w.r.t other miRNAs at the same time and they may lead to loss of information. In this situation, "a weighted framework for integrating fuzzy rough set based relevance and redundancy entropy", called WFIFRRRE, is developed which combines fuzzy set and rough set for ranking and selecting drug resistant miRNAs in cancer.

3.4.1 z score based Entropy Computation

Now we discuss about the way of entropy computation which will be used for miRNA ranking. First we define the fuzzy relevance membership and fuzzy redundancy membership functions using ratio of z scores of an expression value from one class to another

class and then we provide the formulation and schematic diagram for the computation of relevance and redundancy entropies.

Let, $\mu_{\chi_i}(g_j)$ represents the fuzzy membership of g_j in the class χ_i where g_j denotes the j^{th} expression value of miRNA g. The fuzzy membership function is as follows:

$$\mu_{\chi_i}(g_j) = \left(1 + \left(\frac{z_i^c}{z_i^d}\right)^2\right)^{-1} \tag{3.5}$$

where z_i^c and z_i^d are the z scores of g_j in control and drug resistant classes, respectively. z score is a statistical measure to find the distance of data point (g_j) from the group mean in terms of standard deviations. Lower the value of the z score nearer the data point will be to the mean. z score can be calculated as:

$$z_{score} = \frac{g_j - m}{\sigma} \tag{3.6}$$

where, m and σ are the mean and standard deviation of the group. Now, the membership function can be written as:

$$\mu_{\chi_i}(g_j) = \left(1 + \left(\frac{\sigma_i^d}{\sigma_i^c} * \frac{m_i^c - g_j}{m_i^d - g_j}\right)^2\right)^{-1}. \tag{3.7}$$

Here, m_i^c and σ_i^c represent the average and the standard deviation of expression values, respectively, in the control class, and m_i^d and σ_i^d are the average and the standard deviation of the values, respectively, in drug resistant class. Here, $j = 1, 2, \ldots \varrho$, where ϱ represents the number of elements/miRNAs in dataset. The nearness of g_j to the mean of the expression values of the class will result in a higher fuzzy membership of g_j. If g_j becomes equal to the mean of both classes (m_i^c and m_i^d) then $(m_i^c - g_j)$ and $(m_i^d - g_j)$ will result in $\frac{0}{0}$ and it will be treated as 1 as $\lim_{n=0} \frac{n}{n} = 1$ [162].

To compute relevance entropy, the dataset is divided into two classes as control and drug resistant based on patient labels and fuzzy relevance membership of an expression value of a miRNA in a class is computed using Eq. 3.7. On the other hand, for redundancy entropy computation, the whole expression profiles of a miRNA from control and drug resistant classes are considered to be in one class and the expression profile of any the other miRNA is treated as another class. Hence, fuzzy redundancy membership of the expression profile of miRNA can also be computed using Eq. 3.7. This is repeated iteratively for all the miRNAs and the average of membership values is considered as average fuzzy redundancy membership and it is computed as:

$$\mu_{\chi_i}(\tau_j) = \frac{1}{\kappa - 1} \sum_{g_j=1}^{\kappa-1} \mu_{\chi_i}(g_j). \tag{3.8}$$

Here, $\mu_{\chi_i}(\tau_j)$ depicts the average redundant membership of j^{th} miRNA with respect to the other miRNAs and κ represents the number of miRNAs in the dataset.

The boundary region of the fuzzy rough approximation region is $< \overline{\beta}\chi - \underline{\beta}\chi >$. The granules in fuzzy rough lower and upper regions $< \underline{\beta}\chi, \overline{\beta}\chi >$ are determined as:

$$\mu_{\underline{\beta}\chi}(c_i) = \inf_{g} \quad \{max\{(1 - \mu_{c_i}(g_j)), \mu_{\chi}(g_j)\}\}, \quad \forall i, \ and \tag{3.9}$$

$$\mu_{\overline{\beta}\chi}(c_i) = \sup_{g} \quad \{min\{(\mu_{c_i}(g_j)), \mu_{\chi}(g_j)\}\}, \quad \forall i. \tag{3.10}$$

Here, c_i is a fuzzy equivalence class inside the set U/β, g_j denotes the j^{th} value of feature g and χ is a crisp set within universal set U. $\mu_{\chi}(g_j)$ represents the membership of element g_j in set χ where $\mu_{\chi}(g_j) \in \{0, 1\}$.

The membership of g_j in lower and upper approximation set is defined as follows [163]:

$$\mu_{\underline{\beta}\chi}(g_j) = \sup_{c_i \in U/\beta} \quad min\{\mu_{c_i}(g_j), \mu_{\underline{\beta}\chi}(c_i)\}, \ and \tag{3.11}$$

$$\mu_{\overline{\beta}\chi}(g_j) = \sup_{c_i \in U/\beta} \quad min\{\mu_{c_i}(g_j), \mu_{\overline{\beta}\chi}(c_i)\}. \tag{3.12}$$

The cardinality of a crisp set is the sum of the number of objects belonging to the crisp set. The cardinality of a fuzzy rough set can be computed as a sum of the non-zero memberships of miRNAs belonging to the fuzzy rough set. The cardinality of fuzzy rough regions can be calculated as $< \sum_{j=1}^{\kappa} \mu_{\underline{\beta}\chi}(g_j), \sum_{j=1}^{\kappa} \mu_{\overline{\beta}\chi}(g_j) >$ where $g_j \in \chi_i$, and κ is the number of elements in U.

Relative frequency: Average relative frequency is the representation of how frequently miRNA expressions occur in an approximation region. The average relative frequency of each miRNA expression is computed as [161]:

$$\gamma_{rel_{low}} = \frac{1}{i} \sum_{i=1}^{i} \left(\frac{\sum_{j=1}^{\varrho} \mu_{\underline{\beta}\chi}(g_j)}{\varrho_i} \right), \ and \tag{3.13}$$

$$\gamma_{red_{low}} = \frac{1}{\kappa - 1} \sum_{i=1}^{\kappa - 1} \left(\frac{\sum_{j=1}^{\varrho} \mu_{\underline{\beta}\chi}(g_j)}{\varrho_i} \right). \tag{3.14}$$

Here, $\gamma_{rel_{low}}$ and $\gamma_{red_{low}}$ represent the average relevance relative frequency and average redundancy relative frequency, respectively, of lower approximation set. In Eq. 3.13, $i = 2$, as there are only two classes (control and drug resistant). In Eq. 3.14, κ is the number of elements/miRNAs in the dataset as each miRNA is treated as one class for redundancy calculation. ϱ_i is the number of objects having non-zero membership. The

average relative frequency in the boundary region is calculated as [161]:

$$\gamma_{rel_{boun}} = \frac{1}{i}\sum_{i=1}^{i}\left(\frac{\varrho_i - \sum_{j=1}^{\varrho}\mu_{\underline{\beta}\chi}(g_j)}{\varrho_i}\right), \ and \tag{3.15}$$

$$\gamma_{red_{boun}} = \frac{1}{\kappa-1}\sum_{i=1}^{\kappa-1}\left(\frac{\varrho_i - \sum_{j=1}^{\varrho}\mu_{\underline{\beta}\chi}(g_j)}{\varrho_i}\right). \tag{3.16}$$

Now, the relevance ($\mathcal{H}_{rel}$) and redundancy ($\mathcal{H}_{red}$) entropy measures are definded as follows [36]:

$$\mathcal{H}_{rel} = -\frac{1}{2}(\gamma_{rel_{low}} * log_2\gamma_{rel_{low}} + \gamma_{rel_{boun}} * log_2\gamma_{rel_{boun}}), \ and \tag{3.17}$$

$$\mathcal{H}_{red} = -\frac{1}{2}(\gamma_{red_{low}} * log_2\gamma_{red_{low}} + \gamma_{red_{boun}} * log_2\gamma_{red_{boun}}). \tag{3.18}$$

Where, $\gamma_{rel_{low}}$, $\gamma_{rel_{boun}}$, $\gamma_{red_{low}}$, and $\gamma_{red_{boun}}$ are defined in Eqs. 3.13, 3.15, 3.14, and 3.16, respectively.

The entropy computation is mainly divided into six steps:

(a) Categorize the dataset into two groups (χ_1 and χ_2) based on the patients' labels (control or drug resistant),

(b) Calculate the fuzzy relevance membership ($\mu_{\chi_i}(g_j)$) of a miRNA to the control and drug resistant class and fuzzy redundancy membership ($\mu_{\chi_i}(\tau_j)$) of a miRNA w.r.t other miRNAs using Eq. 3.7 and 3.8, respectively.

(c) Determine the membership of a miRNA in fuzzy rough lower and upper regions as $\mu_{\underline{\beta}\chi}(g_j)$ and $\mu_{\overline{\beta}\chi}(g_j)$ using Eq. 3.11 and 3.12, respectively, using fuzzy rough set.

(d) Compute the average relevance ($\gamma_{rel_{low}}$) and average redundancy ($\gamma_{red_{low}}$) frequency of each miRNA in the lower approximation region of fuzzy rough set using Eq. 3.13 and 3.14, respectively.

(e) Calculate the average relevance ($\gamma_{rel_{boun}}$) and average redundancy ($\gamma_{red_{boun}}$) frequency of each miRNA in the boundary approximation region of fuzzy rough set using Eq. 3.15 and 3.16, respectively.

(f) Find the redundancy ($\mathcal{H}_{red}$) and relevance ($\mathcal{H}_{rel}$) entropy of each miRNA using Eq. 3.18 and 3.17, respectively.

The schematic diagram for entropy computation is shown in Fig. 3.1.

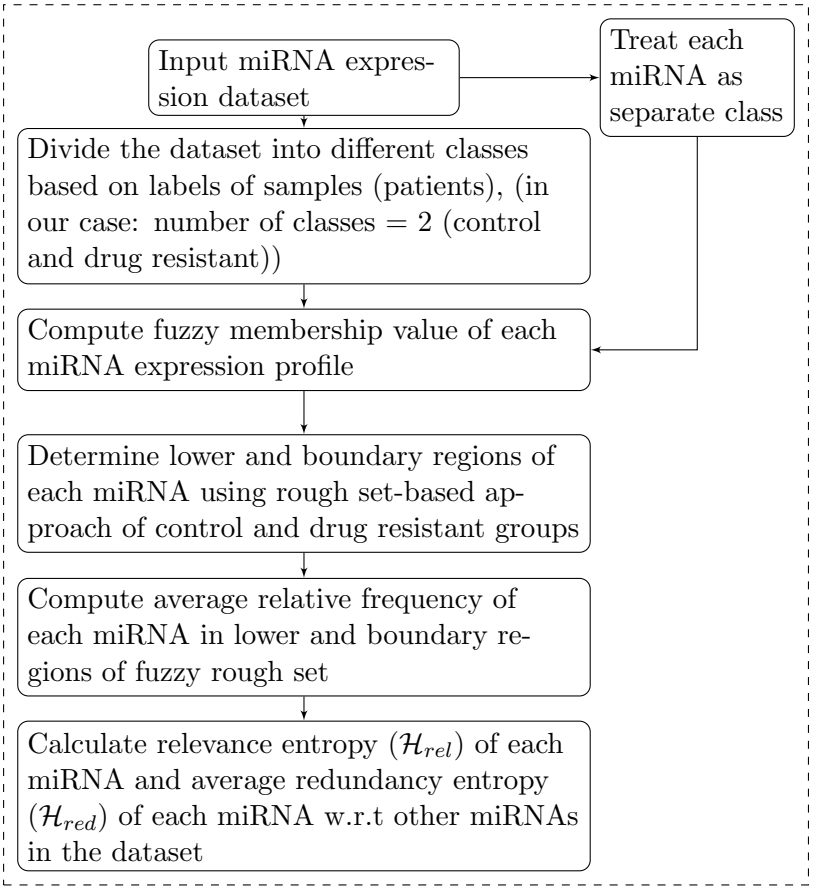


Figure 3.1: Schematic diagram for computation of relevance and average redundancy entropy of each miRNA

3.4.2 Integrating Entropies

The integrated entropy, called WFIFRRRE, is computed as follows:

$$E = w \times \mathcal{H}_{rel} + (1 - w) \times (1 - \mathcal{H}_{red}). \tag{3.19}$$

Here w is varied from 0 to 1 in steps of 0.01 [164, 165] to maximize the classification accuracy of patients in terms of F score. The steps of the developed framework, WFIFRRRE, are as follows:

S1) Initialize the weight w as 0.

S2) Integrate the relevance and redundancy entropies using Eq. 3.19.

S3) Sort the miRNAs in ascending order based on integrated entropy E and select a top percentage of miRNAs.

S4) Classify the patients using selected miRNAs using SVM classifier and compute the F score.

S5) Vary the weight w from 0 to 1 in steps of 0.01 and repeat the Steps S2-S4.

S6) Find the weight w for which the F score is maximum and select those miRNAs as the final set.

The source code for WFIFRRRE and the steps to run the code are provided in the webpage `https://www.isical.ac.in/~shubhra/WFIFRRRE.html` as well as on Github page `https://github.com/joginder12/WFIFRRRE`.

3.5 Experimental Evaluations

WFIFRRRE is also implemented in the Python programming language, and the same machine as EDWFC and HCEDFCR is used to run the code. The performance of WFIFRRRE and the related methods for comparison is evaluated using Naive Bayes (NB), random forest (RF), and linear SVM classifier. The Scikit-learn package [140] is installed in a Python environment for using the classifiers with their default parameters, which are also mentioned in [140]. For example, in NB, the prior probability and variable smoothing are selected as none and $1e^{-09}$, respectively, by default. The command is NB(priors=None, var_smoothing=$1e^{-09}$). For RF, the default values for the number of trees in the forest, the criterion for split, the minimum number of samples needed for split, and maximum features are 100, gini entropy, 2, and the square root of the actual number of features, respectively. The command for RF is RandomForestClassifier(n_estimators=100, criterion='gini', min_samples_split- =2, max_features='sqrt'). The SVM classifier is used with a linear kernel, l2 regularization, squared hinge loss, and a maximum of 1000 iterations. The related command is svm.LinearSVC(penalty='l2', loss='squared_hinge', max_iter=1000). These parameters are also kept the same for all the compared methods for a fair comparison. Leave-one-out cross-validation technique is used for patients to train and test the selected set of miRNAs. The top 1% of miRNAs are chosen for all the compared methods to classify the control and drug resistant patients. The methods are evaluated in terms of sensitivity, specificity, accuracy, F score, and MCC.

3.5.1 Performance Evaluation

The performance of WFIFRRRE is assessed on top 1% of selected miRNAs and those miRNAs are shown in Table 3.2. These miRNAs are further validated using previously existing biochemical/biological studies and discussed in Section 3.6. The classification performance of top 1% miRNAs in terms of F score is shown in Table 3.3. The results in terms of minimum and maximum values of F scores are 0.77 and 1.0, 0.75 and 1.0, and 0.74 and 1.0 for SVM, Naive Bayes (NB), and random forest (RF) classifiers, respectively, for different datasets. On the other hand, using all the miRNAs, these values are 0.19

and 0.62, 0.14 and 0.74, and 0.18 and 0.71. These are shown in Table 3.3. It is clear from Table 3.3 that F score is improved after using top 1% of miRNAs as compared to total number of miRNAs. Similar results are obtained in terms of specificity, sensitivity, accuracy, and MCC and shown in Table 3.4.

Table 3.2: MiRNAs selected by WFIFRRRE from different drug resistant cancer data sets.

Esophageal	Ovarian	Colon FU
hsa-miR-30e-st	miR-324-3p	miR-485-5p
miR-125a-5p-st	miR-149	miR-488
miR-221-st	miR-940	miR-556-3p
miR-1265-st	miR-615-3p	miR-651
miR-27a-st	miR-513a-3p	miR-376a
miR-191-st	miR-20b*	miR-124*
miR-20b-st	miR-150	miR-767-3p
miR-224-st		
Breast	Colon M	Squamous
miR-499-3p	miR-505	miR-99b
miR-21	miR-224	miR-18a
miR-1306	miR-149	miR-125a-5p
miR-26b	miR-193b	miR-1226
miR-320b	miR-505*	miR-935
miR-1305	miR-770-5p	miR-297
	miR-365	miR-146a
Lung	Lymphoblastic	
miR-410	miR-100	
miR-27a	miR-99a	
miR-483-5p	miR-125b	

Table 3.3: Comparing the F score of top 1% miRNAs selected by WFIFRRRE with all miRNAs using SVM, Naive Bayes (NB), and random forest (RF) classifiers.

Cancer Type	F score for top 1% of miRNAs using			F score for 100% of miRNAs using		
	SVM	NB	RF	SVM	NB	RF
Ovarian	0.77	0.75	0.76	0.22	0.37	0.38
Colon FU	0.84	0.82	0.87	0.48	0.54	0.49
Colon M	0.98	0.93	0.96	0.25	0.14	0.18
Esophageal	1.0	0.99	1.0	0.41	0.66	0.44
Lung	0.92	0.85	0.96	0.62	0.34	0.52
Lymphoblastic	0.73	0.75	0.74	0.19	0.33	0.48
squamous	0.90	0.91	0.90	0.50	0.74	0.71
Breast	1.0	1.0	1.0	0.62	0.68	0.61

3.5.2 Comparison with other methods

WFIFRRRE is compared with some well known miRNA and gene selection techniques such as EDWFC [137], SPEM [37], MRMR [33], SVMRFE [38], SVMRFE-MRMR [142], FRMI [36], CBFS [143], FSNLDA [144], FSHDD [145], GSPSO [146], BPSOFS [147], and Lasso [139].

Table 3.4: Classification results for top 1% miRNAs selected by WFIFRRRE using Naive Bayes (NB), random forest (RF), and SVM classifiers of breast, esophageal, lung, and ovarian datasets.

Classifier	Performance Measures	Breast	Esoph-ageal	Lung	Ova-rian	Lympho--blastic	Squa--mous	Colon M	Colon FU
Naive Bayes	Sensitivity	1.0	1.0	0.88	0.76	0.74	0.93	0.86	0.75
	Specificity	1.0	1.0	0.82	0.75	0.76	0.77	1.0	0.93
	Accuracy (%)	100	100	84.37	75.23	75.0	87.5	95.71	83.92
	MCC	1.0	1.0	0.70	0.51	0.50	0.72	0.90	0.70
RF	Sensitivity	1.0	1.0	1.0	0.76	0.71	0.89	0.96	0.93
	Specificity	1.0	1.0	0.92	0.76	0.78	0.80	0.98	0.79
	Accuracy (%)	100	100	95.83	75.55	75.29	86.80	95.56	87.50
	MCC	1.0	1.0	0.92	0.51	0.51	0.67	0.95	0.72
SVM	Sensitivity	1.0	1.0	0.92	0.79	0.72	0.93	0.95	1.0
	Specificity	1.0	1.0	0.92	0.76	0.77	0.71	1.0	0.61
	Accuracy (%)	100	100	91.67	76.05	73.95	85.41	98.57	80.35
	MCC	1.0	1.0	0.83	0.50	0.48	0.67	0.97	0.66

The comparison results are shown in Table 3.5 using top 1% of miRNAs selected by each method. The best outcomes are shown in bold fonts. From the results, it is observed that WFIFRRRE provides the best results in 404 out of 416 cases (8 datasets $\times$ 13 methods $\times$ 4 measures). The F score values for WFIFRRRE are 0.84, 1.0, 0.92, 0.73, 0.77, 1.0, 0.98, and 0.91 for eight datasets. These values are the best and marked in bold fonts in Table 3.5. Interestingly, SPEM and BPSOFS also achieved the best F score (1.00) for the breast dataset. The second best results in terms of F scores for the remaining datasets are also obtained using SPEM and BPSOFS where these values are 0.80 (BPSOFS), 0.87 (SPEM), 0.87 (BPSOFS), 0.66 (SPEM), 0.64 (BPSOFS), 0.94 (BPSOFS) and 0.90 (SPEM).

Similarly, WFIFRRRE also obtained the best results in terms of accuracy for all the 8 datasets where the values are 80.35, 100, 91.67, 73.27, 75.51, 100, 98.57, and 87.50. Note that, SPEM and BPSOFS also achieved the same accuracy (100) for the breast dataset. The second best results in terms of accuracy for the remaining datasets are 76.50 (SPEM), 89.27 (MRMR), 87.00 (BPSOFS), 70.67 (SPEM), 69.25 (SVMRFE-MRMR), 95.00 (BPSOFS), and 85.41 (SPEM).

The sensitivity values achieved by WFIFRRRE are the best for 6 datasets which are 1.0, 1.0, 0.92, 0.79, 1.0, and 0.93 for colon FU, esophageal, lung, ovarian, breast, and squamous datasets respectively. For lymphoblastic and colon M data, SPEM performed better than WFIFRRRE. While using SPEM these values are 0.81 and 1.00, using WFIFRRRE these values are 0.72 and 0.96.

The specificity values achieved by WFIFRRRE are found to be the best for 6 datasets, which are 1.0, 0.92, 0.77, 0.76, 1.0, and 1.0 for esophageal, lung, lymphoblastic, ovarian, breast, and colon M datasets, respectively. WFIFRRRE performed inferior to EDWFC, FRMI, and BPSOFS for squamous, colon FU, and colon FU data, respectively. For

Table 3.5: Comparison of WFIFRRRE with related methods using SVM classifier.

Methods	Performance Measures	Colon FU	Esoph-ageal	Lung	Lympho-blastic	Ovar-ian	Breast	Colon M	Squa-mous
WFIF RRRE	Sensitivity	**1.0**	**1.00**	**0.92**	0.72	**0.79**	**1.00**	0.96	**0.93**
	Specificity	0.61	**1.00**	**0.92**	**0.77**	**0.76**	**1.00**	**1.00**	0.78
	F score	**0.84**	**1.00**	**0.92**	**0.73**	**0.77**	**1.00**	**0.98**	**0.91**
	Accuracy (%)	**80.35**	**100**	**91.67**	**73.95**	**76.05**	**100**	**98.57**	**87.50**
EDWFC	Sensitivity	0.81	0.91	0.75	0.75	0.72	1.00	0.63	0.91
	Specificity	0.71	0.88	0.75	0.71	0.75	1.00	0.81	**0.88**
	F score	0.77	0.89	0.75	0.72	0.74	1.00	0.74	0.89
	Accuracy (%)	76.19	89.58	75	72.99	68.74	100	67.86	**89.58**
SPEM	Sensitivity	0.86	0.83	0.75	**0.81**	0.61	1.00	**1.00**	0.93
	Specificity	0.56	0.92	0.81	0.60	0.72	1.00	0.79	0.71
	F score	0.66	0.87	0.77	0.66	0.59	1.00	0.88	0.90
	Accuracy (%)	62.50	87.50	78.13	70.67	66.07	100	89.28	85.41
MRMR	Sensitivity	0.24	0.87	0.41	0.60	0.59	0.99	0.76	0.71
	Specificity	0.60	0.89	0.32	0.60	0.66	0.99	0.70	0.70
	F score	0.35	0.88	0.36	0.60	0.62	0.99	0.73	0.81
	Accuracy (%)	41.00	89.27	37.48	63.22	62.50	99.99	73.21	70.83
SVM RFE	Sensitivity	0.53	0.38	0.26	0.54	0.57	0.96	0.54	0.63
	Specificity	0.23	0.53	0.57	0.64	0.38	0.91	0.43	0.55
	F score	0.53	0.44	0.34	0.59	0.45	0.94	0.48	0.58
	Accuracy (%)	52.35	44.78	41.67	59.99	47.32	94.25	48.22	58.00
FRMI	Sensitivity	0.63	0.58	0.66	0.65	0.75	0.69	0.81	0.70
	Specificity	**0.76**	0.69	0.78	0.65	0.57	0.59	0.78	0.79
	F score	0.69	0.63	0.71	0.65	0.48	0.63	0.80	0.75
	Accuracy (%)	69.61	63.94	72.66	65.54	62.48	89.78	80.00	75.00
SVM RFE-MRMR	Sensitivity	0.53	0.38	0.24	0.54	0.63	0.96	0.54	0.71
	Specificity	0.23	0.53	0.56	0.64	0.75	0.91	0.43	0.76
	F score	0.53	0.44	0.34	0.59	0.61	0.94	0.48	0.73
	Accuracy (%)	52.35	44.78	41.65	59.99	69.22	94.25	49.34	72.57
CBFS	Sensitivity	0.65	0.53	0.37	0.62	0.63	0.99	0.54	0.54
	Specificity	0.43	0.57	0.43	0.62	0.60	0.99	0.50	0.58
	F score	0.49	0.55	0.39	0.62	0.62	0.99	0.52	0.56
	Accuracy (%)	51.98	56.24	40.62	62.21	61.72	99.99	51.78	56.25
FSN LDA	Sensitivity	0.58	0.78	0.67	0.55	0.56	0.99	0.35	0.79
	Specificity	0.56	0.82	0.48	0.58	0.62	0.84	0.42	0.76
	F score	0.58	0.81	0.57	0.57	0.58	0.90	0.21	0.78
	Accuracy (%)	56.99	80.21	59.36	59.35	59.84	91.99	39.38	78.20
FSHDD	Sensitivity	0.78	0.72	0.71	0.47	0.49	0.83	1.00	0.62
	Specificity	0.60	0.77	0.82	0.67	0.57	0.86	0.72	0.57
	F score	0.71	0.74	0.76	0.55	0.53	0.83	0.84	0.59
	Accuracy (%)	71.74	74.98	76.19	57.57	53.49	85.21	84.57	59.50
GSPSO	Sensitivity	0.65	0.66	0.62	0.44	0.59	0.73	0.45	0.51
	Specificity	0.44	0.57	0.68	0.71	0.61	0.78	0.38	0.65
	F score	0.52	0.61	0.65	0.55	0.60	0.76	0.42	0.56
	Accuracy (%)	54.62	62.02	65.37	58.07	60.73	76.19	38.87	57.19
BPS OFS	Sensitivity	0.90	0.80	0.90	0.63	0.62	1.0	0.96	**0.93**
	Specificity	0.72	0.95	0.84	0.70	0.70	1.0	0.92	0.68
	F score	0.80	0.86	0.87	0.63	0.64	1.0	0.94	0.85
	Accuracy (%)	76.50	87.50	87.00	66.67	68.23	100	95.00	83.50
Lasso	Sensitivity	0.53	0.80	0.65	0.62	0.43	0.96	0.56	0.78
	Specificity	0.25	0.82	0.56	0.58	0.71	0.93	0.45	0.82
	F score	0.54	0.80	0.67	0.59	0.50	0.94	0.51	0.79
	Accuracy (%)	82.42	87.50	62.50	59.35	57.14	95.50	53.75	82.50

Colon Fu, while FRMI performed best (0.76) and BPSOFS performed second (0.72), WFIFRRRE achieved the third (0.61) best result.

In summary, sensitivity, specificity, F score, and accuracy ranges from 0.72 to 1.0, 0.61 to 1.00, 0.73 to 1.00, and 73.95 to 100, respectively, using WFIFRRRE considering all the datasets. In most of the cases, WFIFRRRE outperforms the existing approaches

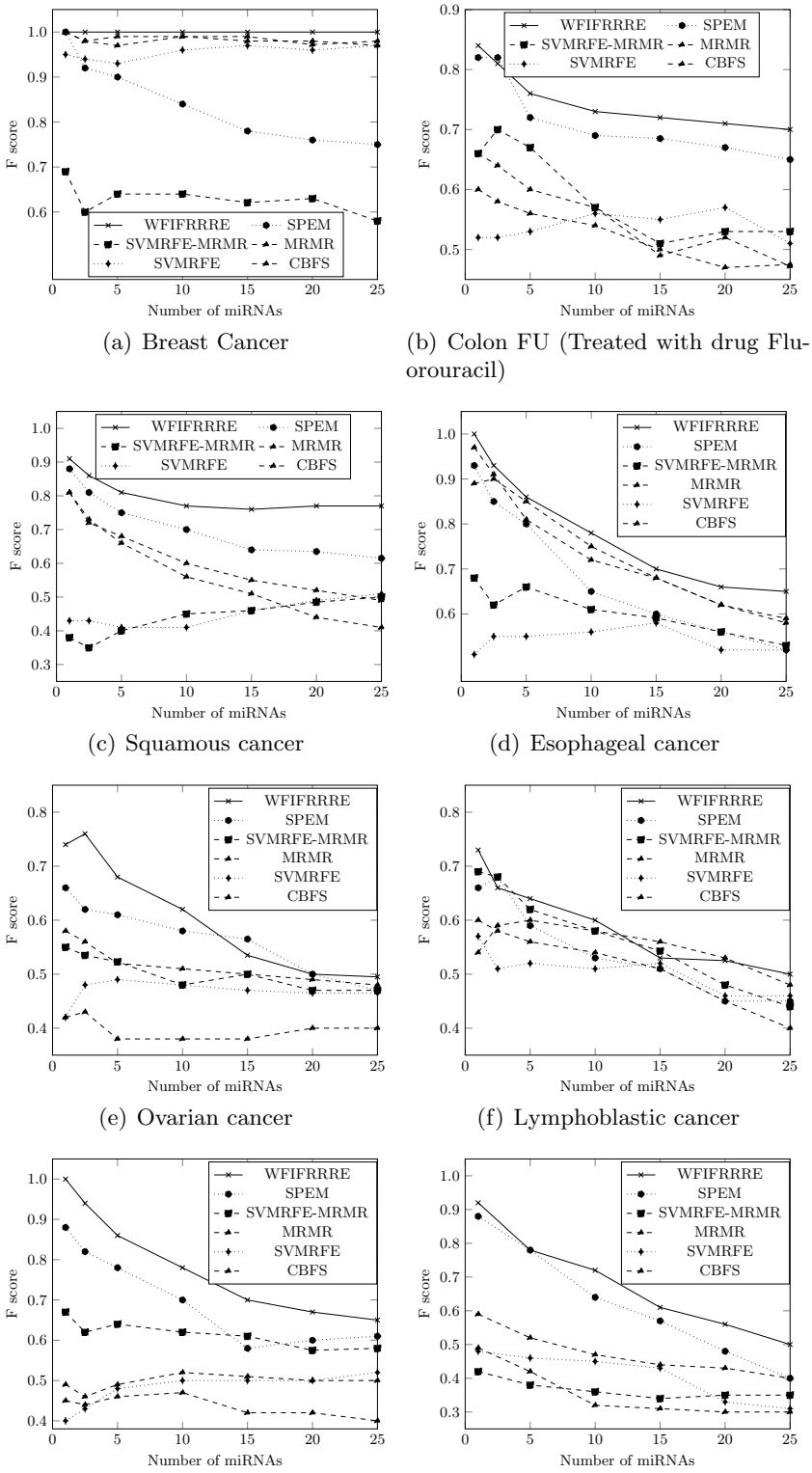
used for comparison.

Table 3.6: Comparison of WFIFRRRE with related methods using RF classifier.

Methods	Performance Measures	Colon FU	Esoph-ageal	Lung	Lympho-blastic	Ovar-ian	Breast	Colon M	Squa-mous
WFIF RRRE	Sensitivity	**1.0**	**1.00**	**0.92**	0.72	**0.79**	**1.00**	0.96	**0.93**
	Specificity	0.61	**1.00**	**0.92**	**0.77**	**0.76**	**1.00**	**1.00**	0.78
	F score	**0.84**	**1.00**	**0.92**	**0.73**	**0.77**	**1.00**	**0.98**	**0.91**
	Accuracy (%)	**80.35**	**100**	**91.67**	**73.95**	**76.05**	**100**	**98.57**	87.50
EDWFC	Sensitivity	0.74	0.90	0.75	0.65	0.68	1.00	0.71	0.90
	Specificity	0.79	0.89	0.75	0.70	0.72	1.00	0.63	**0.89**
	F score	0.74	0.90	0.75	0.67	0.71	1.00	0.65	0.89
	Accuracy (%)	71.88	89.82	75	67.68	66.74	100	66.67	**89.81**
SPEM	Sensitivity	0.79	0.90	0.66	**0.79**	0.63	0.98	**0.96**	0.89
	Specificity	0.56	0.88	0.68	0.64	0.70	0.95	0.80	0.77
	F score	0.66	0.89	0.68	0.67	0.61	0.98	0.89	0.86
	Accuracy (%)	62.50	88.89	67.50	66.20	67.40	96.61	90.45	86.50
MRMR	Sensitivity	0.70	0.83	0.53	0.62	0.63	0.93	0.70	0.86
	Specificity	0.68	0.81	0.50	0.60	0.67	0.97	0.74	0.82
	F score	0.60	0.82	0.52	0.60	0.66	0.95	0.72	0.82
	Accuracy (%)	63.50	81.67	50.71	63.22	66.67	95.50	72.22	85.30
BPS OFS	Sensitivity	0.92	0.81	0.87	0.65	0.64	1.00	0.96	**0.95**
	Specificity	0.75	0.93	0.84	0.71	0.73	1.00	0.92	0.68
	F score	0.82	0.87	0.83	0.65	0.67	1.00	0.94	0.86
	Accuracy (%)	78.50	87.50	84.30	67.50	69.70	100	95.00	84.20

In Table 3.6, the WFIFRRRE is compared with EDWFC, SPEM, MRMR, and BPSOFS using random forest as a classifier. Here, SPEM, MRMR, and BPSOFS are chosen as they lie within the top three methods when using the SVM classifier. From the table, it is clear that WFIFRRRE performs better than the compared methods except for EDWFC using squamous data for specificity (0.88) and accuracy (89.81) measures, for SPEM using lymphoblastic and Colon M data for sensitivity measures (0.79 & 0.96), for BPSOFS using squamous data for the sensitivity measure (0.95). Here, the WFIFRRRE provides the best results in 177 out of 192 cases (8 datasets $\times$ 6 methods $\times$ 4 measures $\times$ 1 classifier). WFIFRRRE performs better than EDWFC in 124 out of 128 cases (8 datasets $\times$ 2 methods $\times$ 4 measures $\times$ 2 classifiers).

Fig. 3.2 shows the variations of F score with different percentages of miRNAs for various methods (WFIFRRRE, SPEM, MRMR, SVMRFE, SVMRFE-MRMR, and CBFS) and datasets. From the curves, it is clear that the F scores of WFIFRRRE are better than the related methods in most of the cases. For example, the curves for WFIFRRRE lie above the related methods for breast, squamous, esophageal, colon M, and lung datasets. The cases where the curves for WFIFRRRE lies below those of related methods are: a) for SPEM using colon FU data in the range 1% to 2% of miRNAs, b) for SPEM using ovarian data in the range 12.5% to 20% of miRNAs, and c) for SPEM & SVMRFE-MRMR using lymphoblastic data in the range 12.5% to 20% & 12.5% to 17.5% of miRNAs, respectively.



(a) Breast Cancer

(b) Colon FU (Treated with drug Fluorouracil)

(c) Squamous cancer

(d) Esophageal cancer

(e) Ovarian cancer

(f) Lymphoblastic cancer

(g) Colon M (Treated with drug Methotrexate)

(h) Lung cancer

Figure 3.2: Comparing WFIFRRRE with related methods in terms of F score for different percentages of miRNAs using SVM classifier.

3.5.3 Complexity of WFIFRRRE

The major components of WFIFRRRE are computation of z-score-based relevance and redundancy entropies, integration of relevance and redundancy entropies using a supervised weight, and classification of patients using selected miRNAs. The complexity discussion for these components is discussed as follows:

Relevance Entropy Computation: Let r and n be the number of miRNAs and patients in a dataset, respectively, where $n = n_c + n_r$ is the total number of samples across the two classes (control n_c and drug resistant n_r). The complexity to compute the class-wise mean and standard deviation of expression values for each miRNA in the z-score (Eq. 3.6) is $O(mn)$. After obtaining fuzzy relevance membership for each sample for every expression value of miRNA, the complexity for determining the lower approximation region, relative frequency, and relevance entropy is also $O(mn)$, $O(mn)$, and $O(mn)$, respectively. The total complexity to compute z-score based relevance entropy is $O(mn + mn + mn + mn) \sim O(mn)$.

Redundancy Entropy Computation: In z-score based redundancy entropy computation in WFIFRRRE, the whole expression profile of a miRNA is treated as one "class" against every other miRNA's profile as the second "class," and then the z-score is calculated. This is repeated for all the miRNAs. If the membership calculated for each miRNA is reused while computing the redundancy w.r.t. other miRNAs, then the complexity involving aggregations across all $m(m-1)$ ordered pairs over n samples for pairwise combinations is $O(m^2 n)$.

Integration of Relevance and Redundancy Entropies: The complexity to integrate relevance and redundancy entropies with a weight is $O(m)$. When the weight, say w, varies from 0 to 1 in steps of 0.01, the complexity to rank the miRNAs at various values of w is $O(w\, m \log m)$.

Patient Classification using selected miRNAs using SVM classifier: The complexity for linear SVM with leave one out cross validation is $O(w\, n\, d)$ [166] where d is the number of selected features. In WFIFRRRE, d is 1% of total miRNAs).

Hence, the computational complexity of WFIFRRRE is

$$O\big(mn \;+\; m^2 n \;+\; w\, m \log m\big) \;+\; O(w, n, d).$$

For high-dimensional miRNA expression data ($m \gg n$ and $m \gg w$), the complexity for redundancy entropy, $O(m^2 n)$, will dominate.

3.6 Biological Evaluation

The biological relevance of miRNAs selected by WFIFRRRE is shown in Table 3.7. Their type of regulation, role in tumour, and target genes are also shown in the table. The role of some of those miRNAs are discussed now.

Lymphoblastic: MiR-125b advances the tumorigenicity of BCR-ABL protein and shows resistance towards vincristine and daunorubicin drug in upregulation with miR-100 and miR-99a [77].

Lung: MiR-483-5p is upregulated in progression of lung adenocarcinoma and it is dysregulated in some other cancers also. MiR-483 is over-expressed in malignant tumor than in benign tumor and it regulates more than 100 genes. Among them 7 genes (ALCAM, FOXJ2, FOX3, RHOGDHI1, NDRG2, SOX11, and TFAP2B) are involved in suppression of cancer metastasis [167]. MiR-410 promotes epithelial-mesenchymal transition (EMT) and radio-resistance in non-small cell lung cancer (NSCLC) cell lines. MiR-410 is upregulated in NSCLC than non-cancerous bronchial epithialial cell lines and promotes cell invasion as well as migration. Hsa-miR-27a shows resistance towards four drugs such as AZD, Erl, Gel, Gö [156].

Ovarian: MiR-324-3p is upregulated by silencing gene AS1 which prevents cell growth. MiR-513-3p is underexpressed in ovarian cancer than other cancers. Over-expression of miR-513a-3p may enhance sensitivity towards the drug cisplatin used in ovarian cancer. [168]. MiR-149 shows sensitivity to cisplatin by suppressing the proliferation in ovarian cancer cell lines. MiR-149-5p expression is upregulated in chemo-resistant ovarian cells. MiR-615-3p suppresses the tumor and its over-expression prevents the proliferation of NSCLC cell lines and reverses the antimatastatic ability of IncRNA HOXA [169].

Esophageal: MiR-20b is downregulated and found to be resistant to drugs in A2780 and A2780/CF70 cell lines. MiR-125a-5p_st is upregulated and increases the cytotoxic effect of cisplatin [170]. MiR-221-3p is sensitive to anti-cancer treatment and dysregulated in different chemotherapeutic-resistant cells. MiR-27a-3p is sensitive to radiation and chemotherapy in esophageal cancer cells. MiR-191-5p is dysregulated in esophageal cancer and shows resistance towards the drug cisplatin [76]. MiR-204 is upregulated and shows sensitivity towards the drug 5-FU. MiR-204 is seen to be down-regulated in many other cancers also.

Colon: MiR-92b-3p shows a higher expression level in tumors of patients who responded to the drug bevacizumab and folfox. MiR-708-5p shows up- and down-regulation in CRC. Its role in prostate and ovarian cancer is also observed. MiR-224 shows resistance towards methotrexate drug resistance, where it is underexpressed [72].

Table 3.7: Biological relevance of the miRNAs selected by WFIFRRRE.

Cancer	miRNAs	Regulation	Role/Function	Target genes
Ovarian	324-3p	Down	Suppressor	WNK2
	940	Down	Suppressor	PKC-δ, SRC
	19a	Up	cell invasion, migration, and proliferation	PTEN
	149	Up	OncomiR	MST1, FOXM1, SAV1
	615	Up	Cell-growth, tumorigenesis and apoptosis	NF-kB2, HOTT1P
	513-3p	Down	chemotherapy Resistance	SMAD4, NFKB1, SMAD3, TP53 and HNF4A
Breast	499-3p	Up	cell apoptosis, proliferation	NBN
	21	Up	control cell proliferation, apoptosis, resistant to Doxorubicine	HER2, CYP1B1, PDCD4, KRAS, NOTCH1, BCL-6
	26b	Down	Suppressor	METTL3, MALAT1, HMGA2
	1305	Up	modifies cancer genes	Not available
Lung	410	Up	metastasis	Not available
	27a	Up	OncomiR, cell proliferation	RXRalpha, BCL-2
	483-5p	Up	Invasion, metastasis	ALCAM, RHOGDI1
Lympho-blastic	100 99a 125b	Up	induced apoptosis, proliferation,	MALAT1, DNTT, SNRPE, KIF5B, NUCKS1, PNO1, RPS11, SET, PRPS2, RPL23A, RPL38
Squa-mous	99b 18a 125a-5p 1226 935	Dys Dys Down Dys Dys	Cisplatin and 5-FU resistant	KRAS ERBB2
Colon M	505	Dys	patients' survival	Not available
	224	Up	suppressor	ADIPOQ, SMAD4, ADH1D, IL1R
	149	Down	OncomiR and suppressor suppressor	FOXM1, SP1, SRPX2, GPC1 EphB3
	193b	Dys	metastasis	Not available
	770-5p	Down	proliferation	HIPK1
	365	Down	Suppressor	Mybl2
Esoph-ageal	30e-st	Down	proliferation, invasion	RPS6KB1
	125a-5p-st	Low	promotes proliferation apoptosis	STAT3, ERBB2
	221-st 27a-st 191-st	Dys Dys Dys	Sensitive to drugs	Not available
	20b-st	Up	metastasis	TP53INP1, RB1
	224-st	Up Up	OncomiR or suppressor	PHLPP1, PHLPP2
Colon FU	485-5p	Down	Suppressor	MCL1
	488	Down	Suppressor	Not available

Breast: Dysregulation of miR-21 is found to increase resistance of CYP1B1, K-RAS, NOTCH1 genes towards the drug doxorubicine (DOX) in breast cancer [71]. Down-regulation of hsa-miR-320b is found to be correlated with lymph node metastasis.

Squamous: MiR-935, 1226, 99b, and 125a-5p are found to be resistant towards the drug 5-FU and sensitive towards chemotherapy in squamous cell carcinoma. 125-5p also shows resistance towards cisplatin in EAC and sensitivity towards chemotherapy. Mir-99b is found to be resistant towards cisplatin in both EAC and ESCC [76].

3.7 Discussion and Conclusion

Two new z score based relevance and redundancy entropies and a way to integrate them in a weighted framework, called WFIFRRRE, for selecting drug resistant miRNAs and classifying patients are presented. The fuzzy membership of each miRNA expression value is computed using the ratio of z scores from one class to the other and these membership values are used to determine the fuzzy rough membership of expression values in lower and boundary regions. The average frequency of expression values of a miRNA in the lower and boundary region of the fuzzy rough set is then determined to compute the relevance and redundancy entropies of each miRNA. Finally, the entropies are integrated through weights in a supervised manner to rank the miRNAs and a portion of miRNAs is selected based on the user's choice. The method, WFIFRRRE, integrates relevance and redundancy of miRNAs in a systematic way to improve the patient classification process. It also judiciously integrates fuzzy logic (FL) and rough set (RS) in the soft computing paradigm. This concept helps not only in handling uncertainty in normal and cancer class overlapping but also in determining the exactness of class size. In soft computing, FL and RS are complementary rather than competitive in nature. Hence, these components are combined in WFIFRRRE for designing a better feature selection technique. This is also a challenging issue in soft computing research.

The classification performance of WFIFRRRE is evaluated using top 1% of the miRNAs on eight datasets. The F score values of classification results, using selected miRNAs, range from 0.73 to 1.0 for various datasets. WFIFRRRE performs better in 404 out of 416 cases while comparing it with other existing methods for various performance measures such as sensitivity, specificity, accuracy, F score, and MCC. Additionally, WFIFRRRE performs better than EDWFC in 124 out of 128 cases. WFIFRRRE also deals with imbalanced datasets, such as esophageal and squamous, where the number of control patients is greater than the drug resistant patients. Hence, WFIFRRRE is more robust than EDWFC. On the other hand, EDWFC is more interpretable than WFIFRRRE as it utilizes the known drug resistant miRNAs from existing literature as biological knowledge to identify the relevant miRNAs. Moreover, the F scores achieved by WFIFRRRE are also found to be better when variable percentages of miRNAs are

used for selection. For 7 out of 8 datasets, the peaks of the curves are observed at 1% of the selected miRNAs. This indicates that not all miRNAs are important in control and drug resistant patient classification and WFIFRRRE can serve the purpose of selecting miRNAs associated with drug resistance in cancer using expression data.

The miRNAs selected by WFIFRRRE are also validated using different existing biochemical/biological studies. Most of the selected miRNAs by WFIFRRRE are mentioned in these investigations. Further, the F score achieved by WFIFRRRE is 0.91 in colon FU where 6 miRNAs are selected. Interestingly, only two miRNAs (485-5p and 488) are found in previous investigations. Hence, the other miRNAs maybe relevant ones in colon cancer and need to be validated in biochemical labs in future. As the findings on different datasets show the importance of WFIFRRRE, it can be also helpful in finding drug resistant miRNAs at different stages of cancer where expressions are available for different stages. Multiclass cancer patients can also be classified using WFIFRRRE by modifying the membership function for multiple class.

Chapter 4

Interpretable Convolutional Neural Network for Selecting miRNAs from Multiple Cancer Classes and Cancer Subtypes through Pan-cancer Analysis

4.1 Introduction

In the previous two chapters, frameworks for identifying drug resistance miRNAs are discussed, and thereafter classifying control and drug resistant patients (two class problem). In this chapter, a methodology is developed for identifying miRNAs for various cancer classes and classifying multiple classes of cancer patients (multiclass problem). The data with various classes of cancer patients is known as pan-cancer data. Note that, at present, there is no drug resistant patient class in pan-cancer data.

Pan-cancer data analysis is about understanding the similarities and dissimilarities of miRNAs or genes in different cancers [171]. The analysis can be carried out using miRNA expression data, which may help in the identification of miRNAs in various cancers. There are certain challenges in designing a computational framework for pan-cancer analysis using miRNA expression data such as some miRNAs may be relevant to more than one cancer class, or some miRNAs may not be relevant at all. Further, the pan-cancer miRNA data is high dimensional and complex (HDC) in nature.

Recently, deep learning models, especially convolutional neural networks (CNNs), gained popularity in HDC data. Hence, to handle HDC pan-cancer data we developed a one dimensional CNN (1D CNN) based framework for pan-cancer analysis using miRNA

expression data. From the existing studies, it is observed that there is no 1D CNN model for miRNA expression analysis which is optimized in layers as well as hyperparameters. Further, there is an existing interpretable 1D CNN model for gene expression [131], but not miRNA expression, that interprets the behavior of the networks but lacks in sensitivity [172] as it uses gradient technique to generate class activation maps. In this situation, we incorporated interpretability in CNN model, called interpretable convolutional neural network model (ICNNM). The novelties of ICNNM are as follows:

i) The number of layers in ICNNM is optimized using Bayesian optimization with a multivariate tree parzen estimator (BoMTPE).

ii) The values of hyperparameters for various layers are also optimized using BoMTPE.

iii) An expected gradient method is used to interpret the developed model which takes care of sensitivity and implementation invariance [173]. The method is used to compute SHapley Additive exPlanations (SHAP) values which provide a good interpretation of the model for identification of miRNAs. For each miRNA, the absolute values of SHAP are computed for test patients and the average is considered as an attribution score for that miRNA.

ICNNM contains a dropout layer at the end of each hidden layer (total 3 hidden layers with one dropout layer in each). This enables every neuron to play a role in solving the task, ensuring that none of them becomes overly specialized in a specific aspect [174]. Note that, till now 1D CNNs are applied for expression analysis without dropout layer. Hence, this addition of a dropout layer in each hidden layer can be considered as an improvement to the existing CNN architectures that are applied for expression analysis only. Once the architecture is fixed, Bayesian optimization with multivariate tree parzen estimator (BoMTPE) is used for optimization of the model. The model optimized in terms of layers as well as other hyperparameters is finally used to classify pan-cancer and cancer subtype miRNA expression data from the TCGA database [88]. Later, an miRNA importance score for ranking miRNAs in each class is developed using Shapely additive explanation (SHAP). SHAP values are computed using expected gradient technique.

The performance of ICNNM is compared with various related CNN models and some popular boosted classifiers using seven datasets. The performance is evaluated in terms of training and test performance. The training performance is determined in terms of train accuracy, train error, validation accuracy, and validation error. The sensitivity, specificity, accuracy, F-score, and MCC measures are used for test performance evaluations. For cross validation, the stratified k-fold validation technique is used. The training performance of ICNNM is compared with related CNN models, whereas test performance is evaluated by comparing ICNNM with related CNN models, pan-cancer classification techniques using miRNA/gene expression data, and some popular well-established boosted classifiers. The ICNNM model performed better than the compared methods in most of the cases.

The chapter is organized in five more sections. Section 4.2 offers a comprehensive summary of datasets used in the study. The developed methodology is outlined in Section 4.3 and Section 4.4 explores the experimental results. Section 4.5 presents a discourse on the efficacy and significance of the results.

4.2 Datasets

The TCGA [1] is a comprehensive database for pan-cancer analysis. The nature of pan-cancer data in TCGA is based on genomics, imaging, and transcriptomics. The TCGA collaborators analyzed the data and developed various computational tools such as the cancer imaging archive (TCIA), the cancer proteome portal (TCTP), Xena, etc. Among them, Xena [88] is a public repository to generate expressions from transcriptomic profiles of miRNA sequences. Here, the expression of a miRNA is represented by $log_2(RPM+1)$ where RPM denotes reads per million mapped reads. The pan-cancer related miRNA sequencing data are generated by Micheal Smith Genome Sciences Center (GSC) and then submitted to the cancer genome atlas (TCGA). The data is reported in raw read counts, reads per million (RPM) and their expressions. The expressions of sequences are submitted in two forms, isomiR and stem-loop, which are vectors of non-zero values carrying relative information. The expressions are represented as a matrix that contains the number of times a sequence successfully aligns with a given reference sequence. For a particular miRNA, the abundance of sequences with respect to the reference sequences provides the expression level.

Xena is used in this investigation to download the data. After downloading the data, miRNAs with more than 20% of missing expressions are removed. For the remaining miRNAs, the K-nearest neighbor imputation [175] method is used to impute the missing expression values. Finally, the miRNA expressions are normalized between 0 and 1 to use them as input for various computational models.

The dataset is described in Table 4.1. It contains 1882 miRNAs and 11,119 samples from 55 classes (33 cancer and 22 normal classes). The number of cancer samples in each class is provided in the third column. In the same column the name of the class is shown in parenthesis. Their full forms are available at the url provided in [1]. There are eleven classes such as ACC, LAML, etc. in the fourth column that contains no normal sample. Hence, those cancer samples whose corresponding normal samples are not available can only be used to distinguish them from other cancer samples. The breast subtype dataset is downloaded from [8]. In Table 4.2, additional information about full forms of the different cancers, the number of normal and cancer patients (sample size) in each cancer, and the sources of each dataset are provided.

The original pan-cancer dataset is divided into 6 datasets and the outline of those datasets is provided in Table 4.3. The classified pan-cancer (CP) data in the second

Table 4.1: Outline of original pan-cancer data. The full forms of the abbreviations of the studies are available in Table 4.2.

Dataset	Number of miRNAs	Number of cancer samples	Number of normal samples	Final data
Original pan-cancer data	1882	79 (ACC) + 413 (BLCA) + 1098 (BRCA) + 309 (CESC) + 36 (CHOL) + 453 (COAD) + 47 (DLBC) + 185 (ESCA) + 5 (GBM) + 525 (HNSC) + 66 (KICH) + 521 (KIRC) + 292 (KIRP) + 188 (LAML) + 530 (LGG) + 375 (LIHC) + 518 (LUAD) + 478 (LUSC) + 184 (PCPG) + 87 (MESO) + 498 (OVAR) + 179 (PAAD) + 499 (PRAD) + 162 (READ) + 263 (SARC) + 450 (SKCM) + 436 (STAD) + 156 (TGCT) + 514 (THCA) + 124 (THYM) = 10,349 (33 classes)	0 (ACC) + 19 (BLCA) + 104 (BRCA) + 3 (CESC) + 9 (CHOL) + 8 (COAD) + 0 (DLBC) + 13 (ESCA) + 0 (GBM) + 44 (HNSC) + 25 (KICH) + 72 (KIRC) + 34 (KIRP) + 0 (LAML) + 0 (LGG) + 50 (LIHC) + 46 (LUAD) + 45 (LUSC) + 0 (MESO) + 0 (OVAR) + 5 (PAAD) + 3 (PCPG) + 52 (PRAD) + 2 (READ) + 0 (SARC) + 2 (SKCM) + 41 (STAD) + 0 (TGCT) + 59 (THCA) + 2 (THYM) =670 (22 classes)	11,119 (55 classes)

Table 4.2: Pan-cancer study type along with their abbreviations, sample sizes, and sources.

Cancer	Abbreviations	No. of samples		References
		Cancer	Normal	
Acute Myeloid Leukemia	LAML	188	0	
Adrenocortical Carcinoma	ACC	79	0	[176]
Bladder Urothelial Carcinoma	BLCA	413	19	[177, 178]
Breast Ductal Carcinoma	BRCA	1098	104	[179]
Cervical Carcinoma	CESC	309	3	[180]
Cholangiocarcinoma	CHOL	36	9	[181]
Colorectal Adenocarcinoma	COAD	453	8	[182]
Esophageal Carcinoma	ESCA	185	13	[183]
Stomach adenocarcinoma	STAD	436	41	[184]
Glioblastoma Multiforme	GBM	5	0	[185, 186]
Head and Neck Squamous Cell Carcinoma	HNSC	525	44	[187]
Hepatocellular Carcinoma	LIHC	375	50	[188]
Chromophobe Renal Cell Carcinoma	KICH	66	25	[182]
Clear Cell Renal Cell Carcinoma	KIRC	521	72	[189]
Papillary Renal Cell Carcinoma	KIRP	292	34	[190]
Lower Grade Glioma	LGG	530	0	[191]
Lung Adenocarcinoma	LUAD	518	46	[192]
Lung Squamous Cell Carcinoma	LUSC	478	45	[193]
Mesothelioma	MESO	87	0	[194]
Ovarian Serous Adenocarcinoma	OVAR	498	0	[195]
Pancreatic Ductal Adenocarcinoma	PAAD	179	5	[196]
Paraganglioma & Pheochromocytoma	PCPG	84	3	[197]
Prostate Adenocarcinoma	PRAD	499	52	[198]
Rectum adenocarcinoma	READ	162	2	[199]
Sarcoma	SARC	263	0	[200]
Skin Cutaneous Melanoma	SKCM	450	2	[201]
Testicular Germ Cell Cancer	TGCT	156	0	[202]
Thymoma	THYM	124	2	[203]
Thyroid Papillary Carcinoma	THCA	514	59	[204]
Uterine Carcinosarcoma	UCS	57	0	[205]
Uterine Corpus Endometrioid Carcinoma	UCEC	542	33	
Uveal Melanoma	UVM	80	0	[206]
Lymphoid Neoplasm Diffuse Large B	DLBC	47	0	

row contains only 33 cancer classes and it is used to classify patients with cancer only. This dataset cannot be used to identify a normal patient from cancer. Hence, to address this issue all normal samples are combined in one class with 33 cancer classes in the classified pan-cancer and normal (CPN) dataset and presented in the third row of Table

4.3. In the fourth row of Table 4.3 the classified normal (CN) dataset is described. Here, only normal samples are used to determine the role of miRNA in normal sample classification. In the fifth row of the table, classified pan-cancer classified normal (CPCN) data is presented where the combination of cancer and normal samples will be used to classify normal and cancer samples for each cancer class. The sixth and seventh row of the table describe the kidney and lung subtype datasets.

The eighth row of Table 4.3 describes the breast subtype data. As mentioned earlier, this dataset is downloaded from [8]. Finally, the ninth and tenth row of the table describes two rare cancer datasets, bone and soft tissue sarcoma (GSE124158) [207] and nasopharyngeal carcinoma staging (GSE32960) [208]. These datasets are downloaded from Gene Expression Omnibus (GEO) [135]. In summary, the ICNNM is evaluated on 9 datasets among which 6 datasets (4 general pan-cancer and two subtypes) are derived from a single TCGA pan-cancer dataset, one dataset is downloaded as Breast sub-class from TCGA, and two datasets, nasopharyngeal carcinoma and bone and soft tissue sarcoma, are downloaded from GEO as rare cancer ones.

The abbreviations of the derived datasets, namely, classified pan-cancer, classified pan-cancer and normal, classified normal, and classified pan-cancer classified normal are CP, CPN, CN, and CPCN, respectively. The same abbreviations are used throughout the article to represent these datasets. In the original pan-cancer data, out of 22 normal classes the number of samples in 6 classes (CESC, PAAD, PCPG, READ, SKCM, THYM) are five or less. Hence, those classes are removed from the derived datasets in Table 4.3. The derived datasets from original pan-cancer data are available at `https://www.isical.ac.in/~shubhra/icnnm.html`.

Table 4.3: Summary of datasets. The datadets in rows 2 to 7 are derived from the original pan-cancer dataset in Table 4.1. The Breast dataset is not a derived one.

Dataset	Number of miRNAs	Number of cancer samples	Number of normal samples	Final data
Classified pan-cancer (CP)	1882	10,349 (33 classes)	0	10349 (33 classes)
Classified pan-cancer and normal (CPN)	1882	10,349 (33 classes)	670 (1 class)	11,119 (34 classes)
Classified normal (CN)	1882	0	670 (22 classes)	653 (16 classes)
Classified pan-cancer classified normal (CPCN)	1882	10,349 (33 classes)	653 (16 classes)	11,002 (49 classes)
Kidney	1882	879 (KICH, KIRC, KIRP)	130 (KICH, KIRC, KIRP)	1009 (6 classes)
Lung	1882	996 (LUAD, LUSC)	91 (LUAD, LUSC)	1087 (4 classes)
Breast	1035	136 (BASAL) + 65 (HER2) + 176 (Luminal B) + 415 (Luminal A) = 792 (4 classes)	25 (normal)	817 (5 classes)

4.3 Interpretable Convolutional Neural Network Model

In the literature, most of the existing CNN models for expression analysis transform the expression into an image and use it as an input to a 2D CNN model. The 2D CNN model requires millions of parameters while training even when expressions can directly be used as an input to a one dimensional CNN. There are few works for expression analysis using 1D CNN [53, 131, 132], but they are not optimized in layers as well as hyperparameters, and the number of layers also varies for different models. Further, a deep learning technique that interprets the behavior of the model is still missing for pan-cancer miRNA expression data. Hence, we develop an optimized 1D CNN model called interpretable convolutional neural network model (ICNNM) to address the aforementioned issues. The developed ICNNM is comprised of three major parts, namely, network architecture, optimization of hyperparameters, and interpretation of the developed model. In the architecture, a dropout layer is added at the end of each hidden layer which removes some neurons with a probability in each step to ensure that none of them becomes overly specialized. After designing the architecture of the CNN model, the number of layers and the other hyperparameters are optimized using Bayesian optimization with a multivariate tree parent estimator (BoMTPE). Later, expected gradient technique is used to compute Shapely additive explanation (SHAP) to interpret the ICNNM. These parts of ICNNM are explained as follows: the network architecture is explained in Section 4.3.1, the optimization of user-defined hyperparameters, including the layers, is discussed in Section 4.3.2, and the interpretation of the developed model is provided in Section 4.3.3.

4.3.1 Network Architecture

The main network of the ICNNM contains 7 kinds of layers. Among those layers, the first three are hidden layers. Each of them also contains four kinds of layers: (i) convolutional layer for extraction of features from the data, (ii) batchnormalization layer for normalizing the distribution of neurons in the convolutional layer, (iii) pooling layer for dimensionality reduction of output from batchnormalization layer, and (iv) dropout layer for eliminating some neurons to avoid overfitting. The remaining three layers among seven kind of layers are: (v) flatten layer to transform the input from previous layer into a single array, (vi) dense layer to generate the feature combination for classification layer, and (vii) softmax layer to find the distribution of data over different classes. Now we discuss about these layers in detail.

Convolutional Layer: Features extraction from the data is performed in the convolution layer. Extracted feature maps which are transmitted to the next layer are obtained using 1D kernel. Feed forward propagation in 1D CNN from previous layer $(l-1)$ to next layer (l) is defined as:

$$\chi_j^k = f\left(\sum_{k \in M_j} \chi_k^{l-1} \times w_{kj}^l + b_j^l\right) \tag{4.1}$$

where χ_j^k and b_j^l represent the j^{th} feature map and bias of l^{th} layer, respectively, w_{kj}^l is the weight between k^{th} neuron at $l-1$ layer and j^{th} neuron at layer l, $\times$ represents convolution operation, χ_k^{l-1} is the feature output from k^{th} neuron at $l-1$ layer, M_j is the size of input feature, and $f(.)$ is the activation function.

Activation Functions: Activation functions are mathematical formulation in neural networks that decide activation/non-activation of neuron in the network. Two activation functions as LeakyReLu and Softmax are used in this architecture where LeakyReLu [209] is used in convolutional and dense layers, and Softmax [210] is used later in output layer as a classifier. The LeakyReLu is defined as:

$$f(x) = \begin{cases} x, & \text{if } x \geq 0, \ \textit{and} \\ \alpha x, & \text{otherwise} \end{cases} \tag{4.2}$$

where x is the input feature to the network.

BatchNormalization layer: Training deep neural networks is difficult due to variations in the distribution of network activations and this behavior is known as internal covariate shift. This problem is addressed by introducing batchnormalization in the network. Batchnormalization reduces the internal covariate shift and accelerates the DNN training by fixing mean and variance of input layer [211].

Pooling Layer: The pooling layer is also known as downsampling layer. It reduces the feature map size and decreases the number of parameters while maintaining translation invariance. The pooling layer can be expressed as:

$$\chi_j^k = f\left(\beta_j^l * downsampling\left(\chi_k^{l-1}\right) + b_j^l\right) \tag{4.3}$$

where β_j^l and b_j^l are the multiplicative and additive biases.

Dropout layer: Dropout is a regularization technique for neural network. In dropout layer, each neuron is kept with a probability p such that it can be dropped with probability $(1-p)$ [212]. Dropout is defined as:

$$r = \kappa * f(Wx) \tag{4.4}$$

where κ denotes the size of binary mask vector with each element drawn independently from $m_j \sim Bernoulli(p)$ and * denotes the element wise product. Bernoulli(p) represent the Bernoulli distribution.

Flatten Layer: The flatten layer receives the output from the last layer and transforms the multi-dimensional output into a single-dimensional vector. This transformation is known as flattening [213].

Dense Layer: Dense layer consists of neurons connected to every neuron from the previous layer. It is also known as fully/completely connected layer. The input is 1 dimensional data and the output is a feature vector. The function for dense layer is defined as:

$$\chi^{ij} = f\left(\sum_{i=1}^{n} W_{ij}^{l-i} * a_i^{(l-1)} + b_j^{l-1}\right) \tag{4.5}$$

where W_{ij}^{l-i} is the weight connecting i^{th} neuron of previous layer and the j^{th} neuron of current layer, $a_i^{(l-1)}$ is the input from i^{th} neuron of previous layer, b_j^{l-1} represent the bias of the j^{th} neuron in the current layer, and $f(.)$ is the activation function.

Softmax layer: This layer computes probabilities of an input feature vector into a class. It is often used in multi-class classification problems where an input can belong to one of several classes [210]. The softmax function is defined as:

$$f(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{c} e^{x_j}} \tag{4.6}$$

where c is the number of output classes.

A loss function is constructed for the network in the developed framework and presented as:

$$loss = CE(y^t, y^p) + \lambda \sum_{j=1}^{c} W_j^2 \tag{4.7}$$

here $CE(y^t, y^p)$ represents the categorical cross entropy between true label (y^t) and predicted label (y^p), and $\lambda \sum_{j=1}^{c} W_j^2$ is the regularization terms in which λ is the hyper-parameter. $CE(y_t, y_p)$ is further extended and the new loss equation is defined as:

$$loss = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{c} (y_{ij}^t * log(y_{ij}^p)) + \lambda \sum_{j=1}^{c} W_j^2 \tag{4.8}$$

where N and c represent the number of instances and classes, respectively.

4.3.2 Optimization of Hyperparameters

Hyperparameters can control the learning parameters of a model [214]. Optimization of hyperparameters can be done automatically through optimization algorithms or manually. Some of the widely known optimization algorithms for hyperparameter tuning are random search, grid search, particle swarm optimization, genetic algorithms, and Bayesian optimization (BO) algorithms [215].

Here, Bayesian optimization using multivariate tree parzen estimator (BoMTPE) is used as a hyperparameter tuning algorithm as it finds the inter-dependencies among hyperparameters if any. BO belongs to the informed search methods where the search algorithm learns parameters from previous iterations to find a better search space [216].

Initially, the number of layers, kernels, filters, batch size, dropout rate, pooling size, and dense size are set to 1, 2, 32, 32, 0.1, 2, and 64, respectively. After that, they are varied in steps of 1, 1, 1, 16, 0.01, 1, and 1 till they reach 6, 5, 256, 128, 0.5, 5, and 256. This is performed using BoMTPE. The BoMTPE finds the optimum values of hyperparameters using a stratified k-fold cross-validation method. The value of k is chosen as 10 and the maximum number of epochs for ICNNM is fixed at 50, as inspired by many studies [217–219], for each fold during training. The objective is to minimize the validation error. An early stopping criterion is implemented to determine the optimal number of epochs by monitoring performance and enhancing the fitness of the model. The model stops if it observes that the validation error has not changed for five consecutive epochs. The optimized hyperparameters are reported in Table 4.4. It can be observed from the table that the optimal number of layers in ICNNM is 3. The batch size (128) and dropout layer rate (0.10) are obtained as the same for all three layers. The filter size and pooling size for the 3 layers in sequence are 121, 107, and 103, and 2, 3, and 2, respectively.

Table 4.4: Optimal hyperparameters for ICNNM.

Layer	Filter Size	Batch Size	Pooling Size	Dropout Rate
Layer1	121	128	2	0.10
Layer2	107	128	3	0.10
Layer3	103	128	2	0.10

4.3.3 Interpretation based Attribution Score

Interpretation of a deep learning model involves examining the model's prediction based on feature (in our case miRNA expression) contribution. Interpretable methods can be categorized into two classes, namely, local interpretation and global interpretation [220]. Local interpretation of a model helps to understand the decision making capability of a model for a single instance [221] while global interpretation models determine the overall behavior of the model [150]. SHapley Additive exPlanations (SHAP) are often

used as global explanation approaches. SHAP uses the coalition game theory concept for a fair distribution of payout among features based on their contributions [222]. The contribution of a feature is calculated as:

$$\chi_j(x) = \sum_{s \subseteq \{x_1, x_2, \ldots, x_m\} \setminus \{j\}} \frac{|s|!(m - |s| - 1)!)}{m!} \times (f_x(s \cup \{x_j\}) - f_x(s)) \tag{4.9}$$

where s represents the subset of features, m is the number of miRNAs to be interpreted, and x is the vector of miRNA expression values to be interpreted. $f_x(s)$ represents the prediction for miRNA expression values in the set s that are marginalized over the set of miRNAs that are not included in set s. $f_x(s)$ can be computed as:

$$f_x(s) = \int \hat{f}(x_1, x_2, \ldots, x_n)\, dP_{x \notin s} - E_\chi(\hat{f}(x)) \tag{4.10}$$

where $\int \hat{f}(x_1, x_2, \ldots, x_n)\, dP_{x \notin s}$ is the multiple integrations performed for each miRNA not contained in s. $E_\chi(\hat{f}(x))$ represents the mean estimation.

This interpretation method helps in finding those miRNAS whose expressions for different classes are responsible for class label prediction. A miRNA attribution score based on expected gradient [173] is computed which follows two of the fundamental axioms sensitivity and implementation invariance [172].

Sensitivity: If two instances say I1 (input) and I2 (baseline) differ in one feature and predict different outcomes then the feature differing in one instance from the other should be given a non-zero contribution value.

Implementation Invariance: Two networks are considered equal if they generate the same result for the same input despite differing implementations.

Expected gradient (EG) is computed due to lack of specific baseline knowledge. Expected gradient can be computed as:

$$EG_i(x) = \int_{x'} ((x_i - x_i^{'}) \times \int_{a=0}^{1} \frac{\partial F(x^{'} + a \times (x - x^{'}))}{\partial x_i} da) \times p_D(x^{'}) dx^{'} \tag{4.11}$$

where $EG_i(x)$ is the EG of input x along with i^{th} dimension, $x^{'}$ is the baseline, and both x, $x^{'} \in R^n$. p_D and a indicate the data distribution and the progress along the integration path, respectively. Integration path is the path along which the inputs are integrated from a baseline to the actual input to compute the attributions. EG, being a path method, fulfils the axiomatic properties of sensitivity and implementation invariance [173].

To interpret the model, a subset of sample is used as an input to the network to compute SHAP values of each miRNA, using Eq. 4.9, that determines their contribution

in class prediction. Based on the SHAP values, a miRNA attribution score is computed using the mean of absolute SHAP value of each miRNA across all samples in each class. Attribution score of a miRNA determines the expression of that miRNAs responsible for that particular class label prediction. It also helps us to rank the miRNAs in a class based on their participation in that particular class.

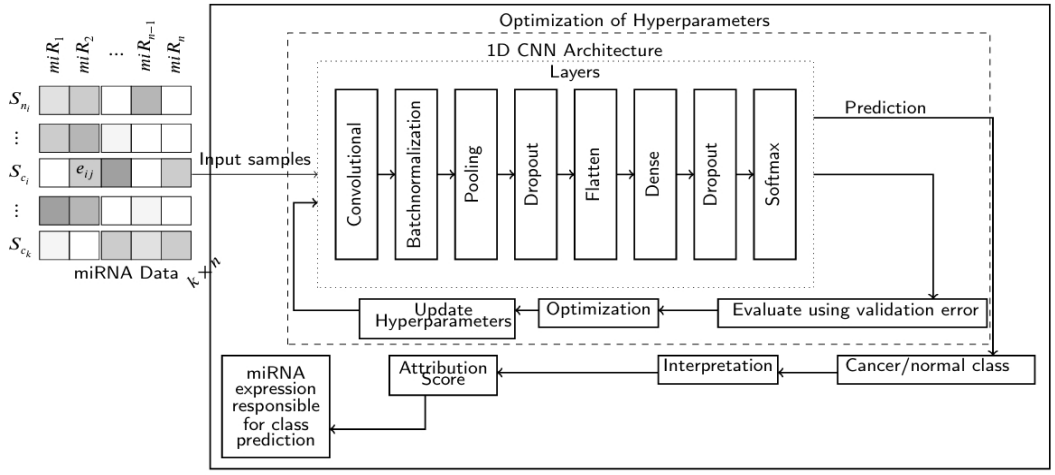


Figure 4.1: Schematic diagram of the ICNNM framework. The dotted rectangle represents the 1D CNN architecture of the developed model and the dashed rectangle shows the part which is optimized (Optimization of hyperparameters). Here, k and n represent the number of samples and number of miRNAs in data, respectively.

A schematic diagram of the ICNNM is shown in Fig. 4.1. It contains the three major parts as discussed above. While the data input is provided in the convolution layer, the output is obtained from the output layer termed as the softmax layer. The source code for ICNNM is provided on Github page `https://github.com/joginder12/ICNNM`.

4.4 Experimental Evaluations

In this section, we discuss the performance and interpretability of ICNNM. In Section 4.4.1, we compare the classification performance of the ICNNM with related CNN models, miRNA or gene selection techniques, and some popular classifiers. Afterwards, to interpret the ICNNM, SHAP values of miRNAs in each class are computed using extended gradients for identifying the responsible miRNAs in class prediction mechanism. The SHAP values are further used to calculate the attribution score of each miRNA in each class as mentioned in Section 4.3.3. The miRNAs with higher attribution scores are considered to be important and selected for further analysis. The analysis of the selected miRNAs is performed in terms of UMAP projection, student t-test, and biological significance in Sections 4.4.4, 4.4.5, and 4.4.6, respectively.

4.4.1 Comparison with Other Methods

To show the efficiency of the ICNNM, the classification performance of the model is compared with several methods. These methods, including ICNNM, are implemented in Python language using Intel(R) Xeon(R) CPU E3-1270 V2 @ 3.50 GHz processor and 32 GB RAM. While the classifiers gradient boosting (Gboost), random forest (RF), and support vector machine with radial basis function (SVM-RBF) are available with the Sklearn package, the Catboost and extreme gradient boosting (XGboost) classifiers are available in packages with the same name. These methods along with their parameters are described as follows:

- One layer CNN model (Base-CNN) [131]: The implemented model consists of one hidden layer and one dense layer. Relu and ADAM are used as activation function and optimizer, respectively. The hyperparameters are number of filters (71), filter size (5), and pool size (2). The optimized values of hyperparameters are obtained using grid search.

- Stacking ensemble-based CNN (Ens-CNN) [53]: The architecture of this CNN model is similar to that in [131] but the hyperparameters are different. The model is implemented as mentioned in [53] where the hyperparameters are number of filters (32), filter size (13), and pool size (2).

- Laplacian and CNN (LS-CNN) based model [132]: In the implementation process, first the Laplacian is used to select the miRNAs, and then the CNN is used for the classification of patients using the selected miRNAs. The CNN is comprised of 6 hidden layers and each layer is comprised of a convolutional layer, a pooling layer, a batch normalization layer, and a relu activation function. Stochastic gradient descent is used as an optimizer. The filter sizes are 1, 8, 16, 32, 64, and 128 for six convolution layers. Max pooling operation is used in the pooling layer. Some additional regularization techniques, such as padding and stride, are also used in each convolution layer.

- GradientBoosting (Gboost) [140]: The parameters used in the command for running the GBoost classifier from SKlearn package are n_estimators=100, learning_rate=0.1, max_depth=3, and random_state=42. Here, n_estimators is the number of trees and max_depth is the depth of each tree while splitting.

- Random Forest (RF) [140]: The parameters of RF from SKlearn package are n_estimators=100, criterion='gini', min_samples_split=2, and random_state=42. Here, min_samples_split indicates the number of splits for a node in a tree. The Gini entropy measure is used as a classification function.

- SVM-RBF: In [119], it is mentioned that the SVM-RBF from Sklearn package is used. The parameters used for SVM-RBF are kernel='rbf', C=1.0, and gamma='scale'.

The C and gamma are parameters for RBF where the gamma parameter implies the range of the influence of a training sample, and C indicates the trade-off between the correct prediction of the training sample and the margin of decision function.

- Catboost [148]: The parameters for Catboost classifier are iterations=1000, learning_rate=0.1, depth=6, and loss_function='MultiClass'. Here the loss function is multiclass cross entropy loss and the depth indicates the depth of each tree.
- XGboost [152]: The parameters for XGboost classifier are objective='multi: softmax', max_depth=3, learning_rate=0.1, and n_estimators=100. Softmax is used as a multiclass classification function.

Seven datasets CP, CPN, CN, CPCN, breast, kidney, and lung are used for comparing different methods. Each dataset is partitioned into train and test sets where 70 percent of the data is used for training, and 30 percent of the data is kept separately for testing. Stratified k-fold (keeping k = 10) cross-validation is performed using this 70 percent of data where out of k folds, k-1 folds of the samples are selected to train the proposed ICNNM model, and the remaining one fold of samples is used for the validation process. After training and validation, the performance of the ICNNM is evaluated with the remaining 30 percent of the test data.

The training performance of ICNNM, Base-CNN, Ens-CNN, and LS-CNN is evaluated in terms of training accuracy, training error, validation accuracy, and validation error. The results are reported in Table 4.5. The best results are obtained for ICNNM for all the 48 cases. For example, the training accuracy, training error, validation accuracy, and validation error for ICNNM using CPCN data are 0.991, 0.061, 0.999, and 0.043, respectively. The second best result is obtained for Ens-CNN.

The training and validation accuracies of ICNNM using CP, CPN, and CPCN data are 0.99±0.05 & 0.998±0.03, 0.993±0.05 & 1.00±0.030, and 0.991±0.03 & 0.999±0.043, respectively. The Ens-CNN is second best performing method which achieves average training and validation accuracy of 0.97 ± 0.08 and 0.96 ± 0.145, respectively.

The variation of training accuracies with epochs for various methods are shown in Fig. 3.2 using CPN data. Six curves are plotted using the training and validation accuracies for Base-CNN, Ens-CNN, and LS-CNN at various epoch values. It can be observed that the gap between training and validation accuracy is minimum for ICNNM and maximum for LS-CNN. For example, the highest gap for ICNNM is 0.05 at epoch 6 and the same for LS-CNN is 0.19 at epoch 13. The averages of the gaps for three methods are also computed using all epochs and they are 0.27, 0.09, and 0.39 for Base-CNN, Ens-CNN, and LS-CNN respectively. It is observed that the average number of epochs for training ICNNM using CPN data is 24 (maximum epochs 50) due to the early stopping criterion.

Table 4.5: Comparing the performance of ICNNM with related CNN models. The bold fonts indicate the best outcomes.

Datasets	Methods	Train accuracy	Train error	Validation accuracy	Validation error
CP	ICNNM	**0.992**	**0.053**	**0.998**	**0.034**
	Base-CNN	0.958	0.150	0.912	0.374
	Ens-CNN	0.971	0.091	0.962	0.123
	LS-CNN	0.950	0.156	0.900	0.477
CPN	ICNNM	**0.993**	**0.052**	**1.00**	**0.030**
	Base-CNN	0.949	0.177	0.900	0.423
	Ens-CNN	0.974	0.086	0.960	0.152
	LS-CNN	0.930	0.225	0.880	0.826
CPCN	ICNNM	**0.991**	**0.061**	**0.999**	**0.043**
	Base-CNN	0.936	0.219	0.894	0.431
	Ens-CNN	0.973	0.087	0.959	0.160
	LS-CNN	0.974	0.085	0.905	0.530

The average number of epochs for CN, CP, CPN, CPCN, breast, lung, and kidney data is observed as 35, 25, 24, 28, 18, 22, and 33, respectively.

Table 4.6: Comparing the test performance of ICNNM with different methods. The bold fonts indicate the best outcomes.

Datasets	Methods	Sensitivity	Specificity	Accuracy	F-score	MCC
CP	ICNNM	**0.96**	**1.0**	**0.99**	**0.97**	**0.96**
	Base-CNN	0.93	0.99	0.98	0.92	0.92
	Ens-CNN	0.92	0.99	0.98	0.92	0.92
	LS-CNN	0.91	0.89	0.90	0.91	0.91
	Gboost	0.78	0.99	0.89	0.88	0.88
	XGboost	0.86	0.99	0.92	0.92	0.92
	Catboost	0.88	0.99	0.93	0.93	0.93
	RF	0.82	0.99	0.91	0.90	0.90
	SVM-RBF	0.82	0.99	0.91	0.90	0.91
CPN	ICNNM	**0.97**	**1.0**	**0.99**	**0.99**	**0.97**
	Base-CNN	0.96	0.99	0.96	0.96	0.94
	Ens-CNN	0.91	0.99	0.99	0.91	0.91
	LS-CNN	0.95	1.0	0.99	0.94	0.94
	Gboost	0.81	0.99	0.88	0.88	0.88
	XGboost	0.89	0.99	0.92	0.92	0.92
	Catboost	0.91	0.99	0.94	0.93	0.93
	RF	0.85	0.99	0.91	0.90	0.90
	SVM-RBF	0.86	0.99	0.91	0.89	0.91
CPCN	ICNNM	0.95	**1.0**	**0.99**	**0.98**	**0.98**
	Base-CNN	**0.96**	0.99	0.99	0.96	0.96
	Ens-CNN	0.88	0.99	0.99	0.95	0.95
	LS-CNN	**0.96**	0.99	0.99	0.96	0.96
	Gboost	0.69	0.99	0.86	0.86	0.85
	XGboost	0.80	0.99	0.92	0.92	0.92
	Catboost	0.83	0.99	0.93	0.93	0.93
	RF	0.77	0.99	0.90	0.89	0.90
	SVM-RBF	0.63	0.99	0.88	0.85	0.88
CN	ICNNM	**0.96**	**1.0**	**1.0**	**0.95**	**0.94**
	Base-CNN	0.95	0.99	0.99	0.93	0.93
	Ens-CNN	0.94	0.99	0.99	0.93	0.93
	LS-CNN	0.75	0.99	0.99	0.71	0.70
	Gboost	0.84	0.99	0.88	0.88	0.88
	XGboost	0.88	0.99	0.92	0.91	0.91
	Catboost	0.85	0.99	0.91	0.90	0.91

Continued on next page

Datasets	Methods	Sensitivity	Specificity	Accuracy	F-score	MCC
	RF	0.83	0.99	0.90	0.89	0.89
	SVM-RBF	0.61	0.98	0.76	0.68	0.75
Breast	ICNNM	**0.90**	**0.98**	**0.98**	**0.91**	**0.90**
	Base-CNN	0.87	0.95	0.93	0.86	0.83
	Ens-CNN	0.85	0.97	0.96	0.87	0.85
	LS-CNN	0.2	0.8	0.79	0.32	0.4
	Gboost	0.56	0.91	0.77	0.75	0.62
	XGboost	0.59	0.92	0.79	0.77	0.67
	Catboost	0.56	0.92	0.79	0.76	0.66
	RF	0.50	0.90	0.76	0.72	0.60
	SVM-RBF	0.47	0.90	0.76	0.70	0.60
Kidney	ICNNM	**0.96**	**0.99**	**0.99**	**0.97**	**0.97**
	Base-CNN	0.96	0.96	0.97	0.96	0.96
	Ens-CNN	0.96	**0.99**	0.99	0.96	0.96
	LS-CNN	0.91	0.9	0.9	0.9	0.88
	Gboost	0.87	0.98	0.94	0.94	0.91
	XGboost	0.93	0.98	0.94	0.94	0.92
	Catboost	0.95	0.98	0.95	0.95	0.93
	RF	0.96	0.98	0.94	0.94	0.91
	SVM-RBF	0.65	0.98	0.90	0.88	0.84
Lung	ICNNM	**0.98**	**1.0**	0.98	0.98	**0.98**
	Base-CNN	0.97	0.98	0.98	0.97	0.95
	Ens-CNN	0.97	0.98	0.98	0.97	0.97
	LS-CNN	0.98	0.99	**0.99**	**0.99**	0.98
	Gboost	0.81	0.99	0.88	0.87	0.87
	XGboost	0.91	0.96	0.92	0.92	0.87
	Catboost	0.93	0.97	0.93	0.93	0.88
	RF	0.92	0.97	0.92	0.92	0.87
	SVM-RBF	0.56	0.94	0.86	0.84	0.76

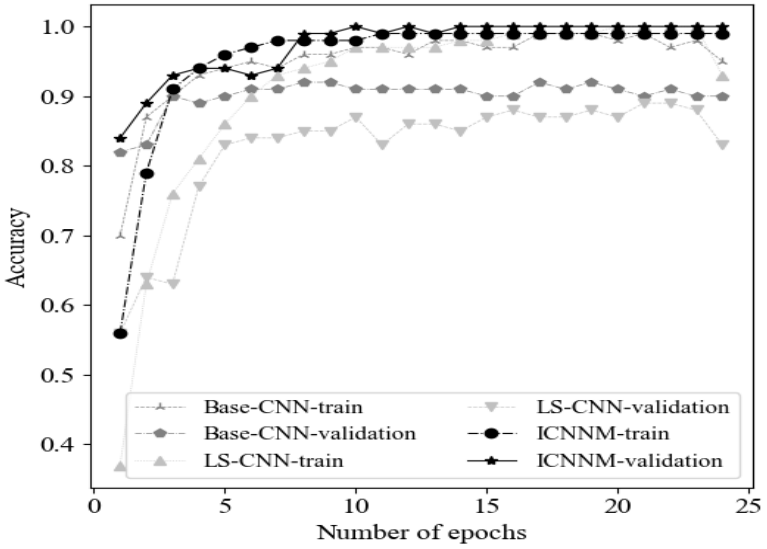


Figure 4.2: Comparing training and validation accuracy curves of ICNNM with LS-CNN and Base-CNN for CPN dataset.

The test performance of ICNNM and related methods in terms of sensitivity, specificity, accuracy, F-score, and MCC is reported in Table 2.4 for CP, CPN, CPCN, CN, breast, kidney, and lung datasets. The fonts in bold show the best results. The F-score values for the ICNNM are 0.97, 0.99, 0.98, 0.95, 0.91, 0.97, and 0.98, for seven datasets.

These values are the best as compared to the related methods except for the lung dataset where LS-CNN achieved the best F-score (0.99). The second best results in terms of F-score for various datasets are obtained using Catboost (0.93), Base-CNN (0.96), Base-CNN and LS-CNN (0.96), Base-CNN and Ens-CNN (0.93), Ens-CNN (0.87), Base-CNN and Ens-CNN (0.96), and ICNNM (0.98).

Similarly, the ICNNM also obtained the best results in terms of accuracy in 6 out of 7 datasets where the values are 0.99, 0.99, 0.98, 0.95, 0.91, and 0.97 for CP, CPN, CPCN, CN, breast, and kidney dataset, respectively. LS-CNN achieved the best accuracy value (0.99) for the lung dataset and ICNNM achieved the second best value (0.98).

The ICNNM also obtained the best results in terms of sensitivity in 6 out of 7 datasets where the values are 0.96, 0.97, 0.96, 0.90, 0.96, and 0.98 for CP, CPN, CN, breast, kidney, and lung, respectively. For CPCN dataset, Base-CNN, and LS-CNN achieved the best sensitivity value (0.96) and ICNNM achieved the second best sensitivity value (0.95). In terms of specificity and MCC, the performance of ICNNM is found to be the best for all 7 datasets.

It can be observed that ICNNM performed better than the related models in 311 cases out of 315 (7 datasets x 9 methods x 5 measures). The 4 cases where ICNNM performed inferior are for Base-CNN and LS-CNN using CPCN data in terms of sensitivity (0.96 and 0.96) and for LS-CNN using Lung cancer data in terms of accuracy and F-score (0.99 and 0.99). In summary, the performance of ICNNM in terms of various measures for general pan-cancer and subtype datasets is as follows:

Sensitivity: The sensitivity of the model varies from 0.95 to 0.97 for general pan-cancer data and from 0.90 to 0.98 for subtype datasets.

Specificity: The specificity values achieved by the model are 1.0 for all of the general pan-cancer datasets, It varies from 0.98 to 0.99 for subtype datasets.

Accuracy: The classification accuracy varies from 0.99 to 1.0 for general pan-cancer datasets and from 0.98 to 0.99 for subtype datasets.

F-score: The F-score achieved by the model varies from 0.95 to 0.99 for general pan-cancer datasets and from 0.91 to 0.99 for subtype datasets.

MCC: The MCC values of ICNNM range from 0.94 to 0.98 for general pan-cancer datasets and 0.90 to 0.98 for subtype datasets.

4.4.2 Interpretability of the Proposed Model

To interpret the ICNNM, SHAP values of miRNAs in each class are computed using extended gradients for identifying the responsible miRNAs in the class prediction mechanism. The SHAP values are further used to calculate the attribution score of each

miRNA in each class as mentioned in Section 4.3.3. The miRNAs with higher attribution scores are considered to be important and selected for further analysis. The analysis of the selected miRNAs is performed in terms of UMAP projection, student t-test, and biological significance in Sections 4.4.4, 4.4.5, and 4.4.6, respectively.

4.4.3 Selection of relevant miRNAs using SHAP values based Attribution Scores

In this section, we describe how the attribution score helps in identifying the relevant miRNAs. To find the relevance of a miRNA in a particular cancer class, the attribution score is computed by considering the average of SHAP values for all the patients (involving that miRNA) in that class. As discussed earlier in Section 4.3.3, SHAP utilizes the game theory concept for assigning scores to the miRNAs based on their performance in prediction of cancer class of patients. A higher attribution score of a miRNA indicates that it is more important than the other miRNAs in class prediction. The SHAP values of a miRNA are obtained in a vector form in Eq. 4.9 where it is defined as the marginalized estimation of miRNA expression values in the set over the expression values of miRNAs that are not included in the set. Finally, the attribution scores for all miRNAs in a particular class are normalized with in 0 to 1 to rank the miRNAs. While an attribution score of 1 in a class means that the relevance is maximum for that miRNA, a score of 0 indicates that relevance of that miRNA is the lowest.

4.4.4 UMAP Projection

Universally manifold approximation projection (UMAP) is a dimensionality reduction technique that helps in data visualization by projecting high dimensional data into a lower dimensional (usually in 2D) space [223]. First, a weighted k-nearest neighbor based fuzzy simplical graph is constructed with the help of a distance metric. Thereafter, a low dimensional graph layout is initialized based on spectral embedding that helps in the formation of an affinity matrix. This matrix is constructed using a similarity function and applying spectral decomposition to graph Laplacian [30]. Finally, approximation is performed by minimizing entropy loss between the initialized layout and the generated k-neighbor weighted graph. The schematic diagram of the UMAP method is provided in Fig. 4.3. For UMAP [223] projection miRNAs having attribution score ≥ 0.2 for each of 33 cancer classes and 1 normal class, using CPN data, are selected. The UMAP is shown in Fig. 4.4. While the classes are overlapping in Fig. 4.4(a) by considering all the miRNAs, the classes are well separated in Fig. 4.4(b) using miRNAs selected by ICNNM. For example, classes 4 & 5, and 3 & 16 are overlapping in Fig. 4.4(a) but those are well separated in Fig. 4.4(b). Similar patterns are also observed for breast, kidney, and lung subtype datasets which are not shown here.

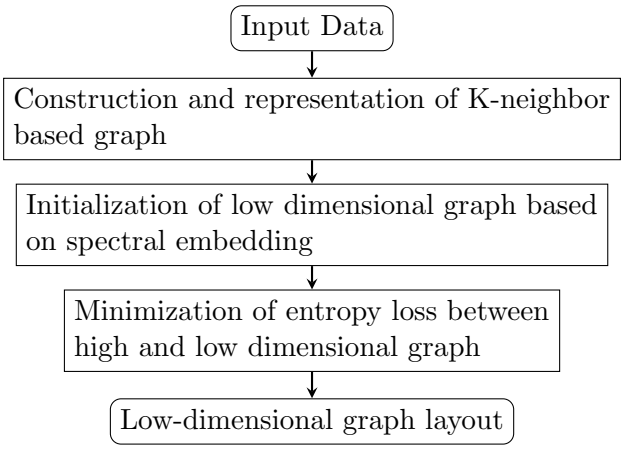
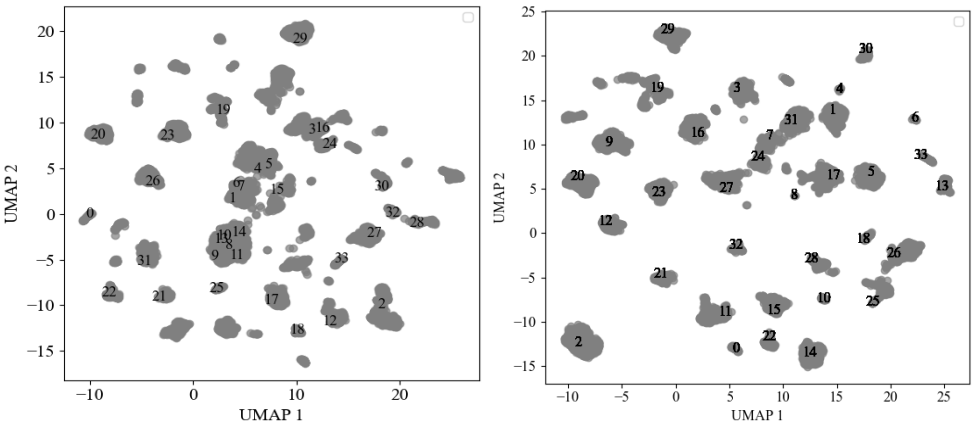


Figure 4.3: Schematic Diagram for UMAP.



(a) (b)

Figure 4.4: Visualization of CPN data using UMAP 2D projection. (a) projection of expressions for all miRNAs and (b) projection of expressions for selected miRNAs.

4.4.5 Discriminability power of Selected miRNAs

The discriminability power of the miRNAs, selected by ICNNM, in differentiating normal and cancer classes is shown by using miRNAs having an attribution score greater than 0.2 in the normal class for the CPN, breast, kidney, and lung data sets. The same miRNAs are also selected from the cancer class along with their expression profiles for each of the datasets. The statistical significance of expressions of normal class with those of cancer class is evaluated using student's t-test. The t value can be computed as follows:

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)}} \tag{4.12}$$

where $\overline{x}$ and $\overline{y}$ denote the means of expression values of normal and cancer classes, respectively, s_x and s_y represent the standard deviations of expression values of normal and cancer classes respectively, and n_x and n_y denote the numbers of samples in normal

and cancer classes, respectively. After computing t statistics, we either accept or reject the null hypothesis that "there is no difference between the expressions of miRNAs in normal and cancer class for a dataset" with a particular significance level based on p-value. The p-value is defined as:

$$p = inf\{\alpha : t > \tau_\alpha\} \tag{4.13}$$

where α is the smallest value at which we reject the null hypothesis and τ_α is the threshold associated with type 1 error. The t-values and the corresponding p-values are provided in Table 4.7. From the results of the t-test for all four datasets, we reject the null hypothesis in favor of the alternative hypothesis that "the expressions of miRNAs are different in normal and cancer class" with a significance level of 0.05.

Table 4.7: Results of t-tests for different pairs of normal and cancer samples using selected miRNAs.

Measures	CPN	Breast	Kidney	Lung
t-value	-3.60	-3.46	17.39	-6.29
p-value	3.07×10^{-4}	2.38×10^{-8}	1.88×10^{-59}	4.39×10^{-10}

4.4.6 Biological Significance of Selected miRNAs

As a computational model, ICNNM selects relevant miRNAs for each class, which helps in the classification of patients. The biological significance of the selected miRNAs is first established with references to existing literature mentioning their role as key biomarkers in respective cancers. Thereafter, the target genes for the selected miRNAs are obtained using the miRDB database [224]. The target genes are then validated using Gene Ontology (GO) [225] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [226]. Note that the selected miRNAs are not biologically validated using any wetlab-based experiments.

The significance of miRNAs selected by ICNNM for each cancer class or subtype is evaluated using attribution score (Eq. 4.9). It represents the relevance of a miRNA to a particular class. The score of a miRNA in each class varies from 0 to 1, where 1 means that the relevance is maximum for that class. Those miRNAs whose attribution score is above 0.50 for a particular class are selected. Fig. 4.5 shows the number of miRNAs having attribution scores greater than 0.5 in each cancer class for the CPN dataset. It is observed from Fig. 4.5 that the number of miRNAs varies from 3 (PCPG class) to 16 (KICH class) for various cancer classes.

While the miRNAs with attribution score > 0.5 for a particular class are available at `www.isical.ac.in./~shubhra/miRNAset_AttSc0.5.pdf`, the miRNAs achieving attribution score > 0.75 are reported in Table 4.8. In the table, cancer classes, miRNAs, attribution scores, target genes, and KEGG pathways are mentioned. The target genes are obtained using miRDB database and validated using gene ontology. In several rows

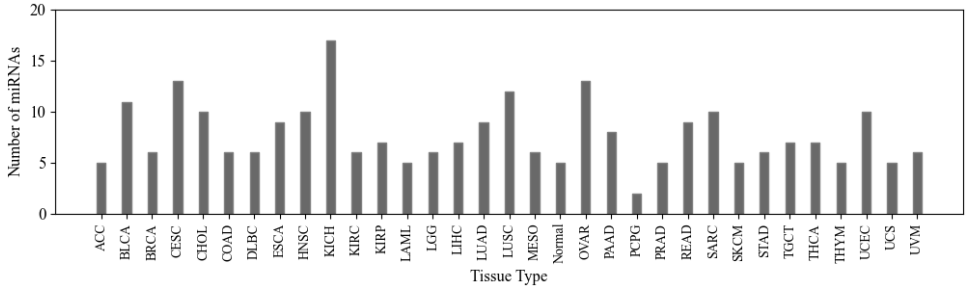


Figure 4.5: Number of miRNAs having attribution score ≥ 0.5 in different classes in CPN data.

of the second column in Table 4.8, it can be observed that more than one cancer class is related to a single miRNA according to the attribution score. Now we discuss those cancer classes and related miRNAs that are obtained using ICNNM and also supported by existing studies.

In the 1^{st} row of Table 4.8, it can be seen that miR-503 targets the genes EPOR, CCND2, MGAT4A, CYP26B1, CCND1, MSH6, ZNF367, MLH1, and PMS2 according to miRDB. The GO analysis of all these genes shows that they are responsible for the positive regulation of isotype switching to IgA (GO:0048298) and IgG (GO:0048304) subtypes, and are also responsible for DNA Repair (GO:0006281). Moreover, it is also mentioned in [171] that four genes, MLH1, MSH2, MSH6, and PMS2, are correlated with cancer class STAD, and EPOR shows a positive correlation with cancer class COAD and a negative correlation with PAAD. Note that miR-503 can be a novel prediction for the cancer classes PRAD and READ (See Table 4.8) as its target genes are related to pathway P13K-Akt in PRAD cancer and pathway MAPK & MSI in colorectal cancer, respectively. It also achieved a high attribution score of 1 for those classes.

Table 4.8: Biological validation of the selected miRNAs by ICNNM. The target genes are obtained using the miRDB database. The cancer names, target genes, KEGG pathway, and related references are mentioned in different columns.

Row No.	Cancer class	MiRNA names	Attribution Score	Target Genes	Gene Ontology	KEGG Pathway	References
1	COAD, ESCA, STAD, PAAD, PRAD, READ	miR-503	1	EPOR, CCND2, CYP26B1, CCND1, MSH6, ZNF367, MLH1, PMS2 MGAT4A, MSH2.	GO:0016446 GO:0048298 GO:0048304 GO:0006281	Pathways in Cancer, MAPK signaling, Insulin signaling, Neurotrophin signaling P13-ATK, MSI	[171]
2	ACC	miR-514a-2	1	NCOA7, ANO5, TMEM68, JAM2, PTPRG, CD200R1	GO:2000401 GO:0030334 GO:2000145	Pathways in Cancer, Ubiquitin mediated proteolysis	[227]
3	BLCA, LIHC, SKCM, CHOL	miR-508	1	PCAT-1, COL4A3BP, ARL4C, UPRT	GO:0009987 GO:0051179	Pathways in Cancer, Focal adhesion, Wnt/ β-catenin-signaling	[228, 229]
4	HNSC, THCA	miR-483	1	IGF2, GAN, PGAM1, AMZ2, AFF1	GO:0043161	Pathways in Cancer, Chronic Myeloid	[230]

Continued on next page

Row No.	Cancer class	MiRNA names	Attribution Score	Target Genes	Gene Ontology	KEGG Pathway	References
						Leukemia, Neurotrophin signaling	
5	SARC, LGG, KIRC, KIRP, MESO, LUAD	miR-202	1	STARD13, ARID3B, CCND2, TRIM71, DICER1, PXDN USP44, MSH2.	GO:0035196 GO:0140374 GO:0000082	Pathways in Cancer, Prostate cancer, Colorectal cancer mismatch repair (MMR)	[231, 232]
6	THYM, BRCA	miR-141	1	QSER1, TMEM170B, ZFR, TNRC6B, EPHA7, TCF12.	GO:0050896 GO:0065007 GO:0009987	Pathways in Cancer, Adherens junction, MAPK signaling	[233, 234]
7	CESC	mir-203a	0.798	ELL2, CAB39, PHIP, MSH2, AFF4, BBX, SIAH1 USP44, THSD7A, ADAMTS6, PDE4D.	GO:0009653	Pathways in Cancer, Ubiquitin proteosomal Wnt signaling.	[235]
8	PCPG	miR-375	1	RLF, POC1B, SPAG9, CENPM, COLCA2, UBE3A ZBTB20, ELAVL4.	GO:0010557 GO:1990090 GO:0031325	Pathways in Cancer, MAPK signaling, Neurotrophin signaling	[236]
9	BRCA	miR-429	0.769	VASH2, MAP2, ZEB1, NR5A2, ERRFI1, SLIT2, HIPK3, WIPF1.	GO:0045765 GO:0030336 GO:0090263	Pathways in Cancer, Focal adhesion, Neurotrophin signaling	[126]
10	LAML, LUSC	miR-10b	1	CADM2, TFAP2C, RORA, CRLF3, GALNT1, E2F7 CNOT6, SOBP.	GO:0045944 GO:2000134 GO:0045893	Pathways in Cancer, Neurotrophin signaling	[237, 238]
11	LGG	mir-10a	0.967	CADM2, CNOT6, , GALNT1, E2F7, CRLF3, SOBP, ELOVL2, TFAP2C . RORA, KCNA6 .	GO:0045893 GO:1902680 GO:0051254	Pathways in Cancer, MAPK signaling, Wnt signaling	[239]
12	DLBC	mir-149	1	CACHD1, ELP5, IFFO2, DLL1, VPS53, KIF2A. STRADB, REPS2 .	GO:0007049	Pathways in Cancer, Endocytosis Neurotrophin signaling	[240]
13	UCS	mir-206	1	HACD3, MMD, CDK14, CPED1, SLC44A1, SMIM14. PAX7.	GO:0065007 GO:0009987 GO:0032502 GO:0008152	Pathways in Cancer, Insulin signaling	[241, 242]
14	KICH	mir-205 mir-30a mir-221	1 0.993 0.921	MOSMO, BICC1, CHN1, CDK19, DCUN1D3, NFAT5, KLHL20, PPARGC18, GABRA1, RIMS3, TCF12, PANK3	GO:0050794 GO:0050789 GO:0045664	Pathways in Cancer, Wnt Signaling, Pathways in Cancer, MAPK signaling Pathways in Cancer, ErBB signaling	[153]
15	UCEC, TCGT, CESC	mir-135a-2	1	SIAH1, ZNF99, CLDN20, ZNF208 ZNF138.	GO:0006357	Pathways in Cancer, MAPK signaling, Wnt signaling	[235, 243]
16	OVAR	mir-200c mir-504	1 0.830	NRP1, VASH2, ZEB1, NR5A2, WDR47, PRR11, VAMP3, UPK1B	GO:0001938 GO:0000226 GO:0045766	Pathways in Cancer, Ubiquitin proteosomal Pathways in Cancer, Chronic Myeloid	[244]

Continued on next page

Row No.	Cancer class	MiRNA names	Attribution Score	Target Genes	Gene Ontology	KEGG Pathway	References
				MAP2, MITF.		Leukemia	
17	SARC	mir-182 mir-183	0.948 0.773	TIAM1, RGS17, PFN2, BNC2, MITF SNX30, MEF2C, GNG5, UNC13B.	GO:0030318 GO:0006355 GO:0065003	Focal adhesion, Mapk signaling, Endocytosis, Neurotrophin signaling,	[245]

In the 3^{rd} row of the table, miR-508 targets the genes PCAT-1, COL4A3BP, ARL4C, and UPRT according to miRDB. These genes take part in cellular localization (GO: 0051179). PCAT-1 is found to be upregulated in LIHC and BLCA [229]. PCAT-1 is also found to be related to cancer CHOL [246]. In [228], miR-508 is also reported as a key biomarker in SKCM.

In the 4^{th} row, miR-483 is observed to target the genes IGF2, GAN, PGAM1, AMZ2, and AFF1. These genes are responsible for proteasome-mediated ubiquitin-dependent protein catabolic process (GO:0043161) as per GO analysis. Further, according to [230], IGF2 is found to be related with cancer classes HNSC and THCA along with other classes.

In the 5^{th} row, miR-202 is found to be related to the cancer classes LGG, KIRC, KIRP, MESO, LUAD, and SARC. In [231], it is reported that mir-4435-2hg can sponge miR-202 and promote cancer progression in kidney renal cell carcinoma (KIRC), Kidney renal papillary cell carcinoma (KIRP), lung carcinoma (LUAD), gliomas (LGG), and osteosarcoma (SARC). Hence, for the remaining class MESO, miR-202 can be considered as a novel prediction as its target gene MSH2 is related to MMR pathwat in MESO [247]. It also achieved a high attribution score of 1 for that class.

In the 14^{th} row, three miRNAs (miR-205, miR-30a, miR-221) are found to be related with KICH cancer [153]. This information is supported by "ENCORI: an encyclopedia of RNA interactome" [153] database where it is mentioned that these miRNAs are differentially expressed in KICH.

In the 15^{th} row, miR-135a is found to be related with genes SIAH1, ZNF99, ZNF138, ZNF208, and CLDN20. The GO analysis of all these genes shows that they are responsible for the regulation of transcription by RNA polymerase II (GO:0006357). According to [235], miR-135a is found to be key biomarker in cervical cancer (CESC) with target gene SIAH1. Further, upregulation of miR-135a is related to proliferation, invasion, and migration of cells in uterine cancer (UCEC) by suppressing MMP2 and MMP9 proteins [243]. Hence, for the remaining class TGCT, miR-135a can be considered as a novel predictionas MAPK pathway is found to be regulated in TGCT [248]. Moreover, according to NCBI (`https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=7697`) the target gene ZNF138 shows ubiquitous expression in testis. MiR-135a also achieved a high attribution score of 1 for TGCT class.

In 16^{th} row, miR-200c and miR-504 are found to be related to ovarian cancer and some of the target genes are NRP1, ZEB1, MAP2, MITF etc. These genes are responsible for the positive regulation of angiogenesis (GO:0006357) and positive regulation of endothelial cell proliferation subtypes (GO:0001938). Further, it is found that miR-200c suppresses gene NRP1 in ovarian cancer that is resistant to therapy and miR-504 is also related with ovarian cancer [244].

In 17^{th} row, one of the target genes, TIAM1, for miR-182 is found to be related with sarcoma cancer (SARC) [245]. Moreover, lower expression levels of miR-182 and miR-183 are found to be correlated with advanced clinical stage, metastasis, and cancer size in SARC [245]. The GO analysis of the target genes shows that they are responsible for the assembly of protein-containing complex (GO:0065003).

For the remaining rows in Table 4.8, the information about miRNAs and the related cancer classes are discussed. In the second row, miR-514a-5p is found to be related to cancer ACC which is also reported in [227]. The role of miR-141 (6^{th}) row) as a key biomarker in BRCA and THYM is established in [233] and [234], respectively. MiR-203a (7^{th} row) is found to be a key biomarker in cervical cancer [235]. MiR-375 (8^{th} row) is found overexpressed by upregulating Yes-associated protein 1 and downregulating axin2 and beta-catenin in pheochromocytoma (PCPG) cancer [236]. MiR-429 (9^{th} row) is reported as associated with homologous recombination deficiency in breast cancer [126]. A deficiency in homologous recombination leads to better survival of patients with breast cancer.

According to GO analysis, regulation of angiogenesis (GO:0045765), negative regulation of cell migration (GO:0030336), and positive regulation of canonical Wnt signaling pathway (GO:0090263) are some of the biological processes associated with the target genes of miR-429.

MiR-10b (10^{th} row) is seen to regulate the apoptosis and proliferation of acute myeloid leukemia (LAML) [237] and lung cancer (LUSC) [238]. The target genes are found responsible for positive regulation of transcription by RNA polymerase II (GO:0045944), positive regulation of DNA-templated transcription (GO:0045893), and negative regulation of G1/S transition of mitotic cell cycle (GO:2000134). In the 11^{th} row, our finding is supported in [239] where miR-10a is reported as a key biomarker in lower grade glioma (LGG). The target genes of miR-10a positively regulate DNA-templated transcription (GO:0045893), RNA biosynthetic process (GO:1902680), and RNA metabolic process (GO:0051254). In the 12^{th} row, miR-149 is found to be related to cancer DLBCL which is also reported in [240]. The target gene is found to be associated with the GO process cell cycle (GO:0007049). In the 13^{th} row, our observation is supported in [242] where miR-206 is found to be differentially expressed in uterine cancer (UCS). The target genes of miR-206 are associated with developmental process (GO:0032502) and metabolic process (GO:0008152).

In summary, 18 out of 21 miRNAs, shown in Table 4.8, are mentioned as key biomarkers in their respective cancer classes according to existing investigations. Three miRNAs, miR-202, miR-503, and miR-135a, reported in Table 4.9 achieved an attribution score of 1 for certain cancer classes but not mentioned as key biomarkers for any cancer class in existing literature. Interestingly, the target genes of these miRNAs are found to be related to cancer-specific pathways as shown in Table 4.8. Hence, miR-503, miR-202, and miR-135a can be considered as novel predictions for cancer classes PRAD & READ, MESO, and TGCT, respectively but further biochemical experiments need to be conducted for confirmation.

Table 4.9: Novel class prediction of miRNAs based on the associations of their target genes in certain cancers.

miRNA	Target Genes	KEGG pathways	Cancer Classes related to pathways	Predicted Classes by ICNNM
miR-503	CCND1	P13-ATK	PRAD	PRAD
	CCND1, MLHI, MSH1, MSH6	MSI, MAPK	Colorectal cancer	READ (Rectum Adenocarcinoma)
miR-202	MSH2	MMR	MESO	MESO
miR-135a	ZNF138	MAPK	TGCT	TGCT

MiRNAs in Subtypes: The top 3 miRNAs (miR-135b, 18a, 190b) found in breast subtype by ICNNM are mentioned as key biomarkers in breast cancer diagnosis in [249]. In the lung subtype, four miRNAs (miR-326, 205, 6510, 375) out of the top 5 miRNAs are found to play a significant role in lung cancer apoptosis and metastasis [250]. The top five miRNAs selected from kidney subtypes based on attribution score are miR-141 [251], 3607 [252], 155 [251], 34a [253], and 204 [254]. These miRNAs are reported in many researches as responsible biomarkers in kidney cancer.

4.4.7 Complexity of ICNNM

There are three major components of ICNNM: 1D CNN, hyperparameter optimization using Bayesian Optimization, and SHAP value computing for identifying relevant miRNAs. Moreover, CNN is the major component of ICNNM, as the hyperparameters of CNN are optimized using BoMTPE, and then the SHAP values are computed using the resultant CNN as a classifier to determine the contribution of miRNAs in the prediction of the patients. Therefore, the complexity is also dependent on these three modules, and it is discussed as follows.

BOTPE: The time complexity for optimizing hyperparameters of a 1D CNN using Bayesian optimization with the tree parzen estimator depends on the number of objective functions, the number of iterations, the complexity of surrogate modeling, the cost of training a 1D CNN, and the number of hyperparameters to be optimized. Suppose the complexity of a 1D CNN model is C_f. TPE scales linearly in terms of past evaluation E and number of hyperparameters H. The time required can be computed in $O(H \times E)$.

1D CNN: The time complexity of 1D CNN consisting of one convolutional layer can be computed in terms of the number of operations required to process a sample (patient) with n features (miRNAs). The number of operations in a convolution layer can be determined by the kernel size k, the number of filters f, and the length of the input vector n. The convolution operation is an elementwise dot product between the kernel and the input vector. For a convolution layer, the number of operations for one kernel is $n \times k$, and considering all the kernels, it is $f \times k \times n$. If the model consists of L layers, then the complexity of the model is $O(L \times f \times k \times n)$. The process is repeated to process all the samples M, iteratively, hence the complexity (C_f) is $O(M \times L \times f \times k \times n)$.

Attribution Score using SHAP: The complexity for SHAP is 2^n where n is the number of features (miRNAs) [255]. In the process, it calculates all the possible combinations of features. In this investigation, SHAP values are computed using the expected gradient methods (Eq. 4.9) where the change in gradient of the original sample with respect to the baseline sample is computed using MOHCNN as a classifier. The integration of SHAP with a classifier does not require retraining of the classifier and it allows fewer computation than 2^n [256]. Hence, the complexity in integrating SHAP with MOHCNN in the class score is dependent on gradient computation for a sample using MOHCNN, the number of points in estimating the gradient (k), and the number of baseline points (S). As mentioned earlier, suppose the complexity of a 1D CNN is C_f. The complexity to explain a sample can be written as $O(S \times k \times C_f)$. For all the samples M, the complexity can be computed as $O(M \times S \times k \times C_f)$.

In summary, the computational complexity of the ICNNM is $O(H \times E) + O(M \times L \times f \times k \times n) + O(2^n) \approx O(H \times E \times C_f) + O(M \times P_D \times C_f) + C_f$.

4.5 Discussion and Conclusion

In this chapter, an interpretable optimized 1D convolution neural network model for pan-cancer and cancer subtype classification, named as ICNNM, is developed. The model determines the optimal hyperparameters for training the network using Bayesian optimization with a multivariate tree Parzen estimator (BoMTPE). Modification in the base CNN model by adding batchnormalization layer and dropout layer in each of the convolution layers is also performed. The batchnormalization layer reduces the change in network parameters during training by setting the mean and variance of the input to a particular value. The dropout layer decreases the overfitting of the complex network while training on small sample datasets such as breast, lung, and kidney subtypes. MiRNA expression data is used in a vectorized format as an input to the model. It is pointed out in several investigations that CNN models with one-dimensional kernels perform well in finding relevant patterns in heterogeneous tensor-like complex biological

data, such as gene expressions and biomedical images [257–259]. This is accomplished through back-propagation along with a gradient descent-based learning process.

The developed model (ICNNM) handles the miRNA expression data with varying sample sizes, such as CPN containing 11,119 samples and the kidney dataset containing 1009 samples. It is observed from Table 4.6, ICNNM outperforms the related methods in terms of sensitivity, specificity, F-score, accuracy, and MCC in 311 out of 315 cases, while comparing with related techniques. The Ens-CNN is observed as the second-best method. The model also achieves lower training and validation errors (Table 4.5) than those of the compared models. For example, the average training and validation errors are 0.034 and 0.01, respectively, for the developed ICNNM which is a one-dimensional CNN model with three layers and five hyperparameters. In contrast, the error values are 0.123 and 0.233 for the second-best method, Ens-CNN, which is an ensemble of 5 one-dimensional 1-layer CNN models.

The interpretation approach of ICNNM, based on SHAP values, shows its significance in ranking the miRNAs based on their contribution to the prediction of class labels of patients. The miRNAs selected by ICNNM also show their discriminability power in distinguishing cancer classes as shown in Section 4.4.5. For example, the selected miRNAs are found significant with p-value <0.05 for all the datasets which show the acceptance of the alternative hypothesis, "the expressions of miRNAs are different in normal and cancer class". Further, some of the selected miRNAs are also found to be biologically significant according to the existing studies. For example, miRNA-141 and 146a are selected from CP dataset and they are pointed out as important biomarkers in various cancers in [233, 234, 260–262] and [263–265], respectively.

Ninety miRNAs (approximately 3 miRNAs from each class, and there are 34 classes) are selected from CPN data based on the attribution score (varies from 0.20 to 1.0). Similarly, the attribution scores of selected miRNAs from breast, kidney, and lung subtype datasets range from 0.7 to 1.0, 0.5 to 1.0, and 0.45 to 0.95, respectively. The average F-scores for these datasets are 0.91, 0.97, and 0.99. Most of the miRNAs selected by ICNNM for various datasets are also mentioned in numerous studies as miRNA biomarkers.

Chapter 5

Set-theoretic Explainable AI-based Attribution Score for Identifying miRNAs in Pan-cancer Data

5.1 Overview

In Chapter 4, a 1D convolutional neural network (1D CNN) is optimized in single objective framework for identifying miRNAs in various cancer classes of pan-cancer data. The use of SHAP values to interpret the relevance of miRNAs across multiple cancer classes is also demonstrated. In this chapter, a multi-objective framework for optimizing hyperparameters of a 1D CNN, called MOHCNN, and a set-theoretic explainable AI-based attribution scores (STEAAS) for miRNA selection are developed. The same dataset as in Chapter 4 is used. The contributions are as follows:

1. 1D CNN is optimized in a multi-objective framework in terms of layers as well as other hyperparameters.
2. A new set-theoretic explainable AI-based score is defined for each miRNA, which has two parts
   (a) determining the importance of the miRNA belonging to a particular cancer class as a class score, and
   (b) the level of reliability of the miRNA belonging to a particular cancer class.

   Therefore, every miRNA has scores for all the cancer classes.

As mentioned earlier in Section 1.3.1.3, most of the existing CNN models for analyzing expression data deal with genes rather than miRNAs (See Section 1.3.1.3). In [131] and [128], expression data is used, and most of the hyperparameters are optimized but the number of layers is fixed. Hence, in our previous work, described in Chapter 4, the number of layers and other hyperparameters are optimized in a single objective framework using miRNA expression data. In this chapter, the hyperparameter optimization of a 1D CNN is performed in a multi-objective framework (MOF) for the same data. Note that optimization of 1D CNN in the multi-objective framework is not performed for expression datasets to the best of our knowledge. The MOF is used to find a trade-off between the training, validation, and number of training parameters of the CNN model. The method is called optimized 1D CNN (MOHCNN). Moreover, a new set-theoretic explainable AI-based attribution score (STEAAS) is also defined, which uses the $Z - number$ concept, to identify the relevant miRNAs. In (STEAAS), a miRNA with a high class score and a high reliability score for a class is treated as a relevant miRNA in that cancer class. A user-defined threshold value of 0.7 is chosen, for both class score and reliability score, above which a miRNA is considered as a relevant one. The selected miRNAs are then validated using publicly available databases such as the ENCORI/Starbase and OncomiR. The regulated genes of the corresponding miRNAs in different cancers are obtained from these databases and also reported in this study.

The performance of MOHCNN is evaluated and compared with three related CNN models, three methods developed for pan-cancer patient classification using miRNA expression data, and two highly cited classifiers using seven datasets. All the datasets are partitioned into train and test data for determining the training and test performance of the model. Sensitivity, specificity, accuracy, F-score, and MCC (See Section 1.3 for these measures) measures are used to assess the test performance of the model. Stratified K-fold CV is used as a CV technique for model training. The MOHCNN model outperforms the compared methods in most of the cases. It is found that most of the miRNAs from the selected ones are found to be key biomarkers in various cancer classes.

The chapter is arranged as follows: an outline of the datasets is provided in Section 5.2, the developed framework, MOHCNN, is explained in Section 5.3, and Section 5.4 is dedicated to experimental results and performance comparison of the framework. Finally, the key findings and results are discussed and concluded in Section 5.6.

5.2 Datasets

The datasets used in this investigation are the same as those mentioned in Section 4.2 of Chapter 4. The summary of the datasets is also provided in Table 5.1 for the convenience of the readers. The details of the subtypes for a cancer class, mentioned in parenthesis of column three of the table, are also mentioned in Section 4.2 of Chapter 4. Note that,

hyperparameters of MOHCNN is optimized with CPCN data with 33 cancer classes and 16 normal classes. As mentioned in Chapter 4, this data set is derived from the original pan-cancer data by removing 6 normal classes where the number of normal patients is less than 5. This is performed to avoid biasness in the optimization of MOHCNN.

Table 5.1: Summary of datasets. The datasets in rows 2 to 7 are derived from the original pan-cancer dataset as mentioned in Chapter 4. The Breast dataset is not a derived one.

Dataset	Number of miRNAs	Number of cancer samples	Number of normal samples	Final data
Classified pan-cancer (CP)	1882	10,349 (33 classes)	0	10349 (33 classes)
Classified pan-cancer and normal (CPN)	1882	10,349 (33 classes)	670 (1 class)	11,119 (34 classes)
Classified normal (CN)	1882	0	670 (22 classes)	653 (16 classes)
Classified pan-cancer classified normal (CPCN)	1882	10,349 (33 classes)	653 (16 classes)	11,002 (49 classes)
Kidney	1882	879 (KICH, KIRC, KIRP)	130 (KICH, KIRC, KIRP)	1009 (6 classes)
Lung	1882	996 (LUAD, LUSC)	91 (LUAD, LUSC)	1087 (4 classes)
Breast	1035	136 (BASAL) + 65 (HER2) + 176 (Luminal B) + 415 (Luminal A) = 792 (4 classes)	25 (normal)	817 (5 classes)

5.3 Methods

In MOHCNN, a 1D CNN is first implemented and then a multi-objective framework using Bayesian optimization is developed for optimizing its hyperparameters. A set-theoretic explainable AI-based attribution scores is also formulated for selecting relevant miRNAs. The network architecture of conventional 1D CNN and the developed MOHCNN are described in Sections 5.3.1 and 5.3.2, respectively. The hyperparameter optimization process in multi-objective framework is discussed in Section 5.3.3. Finally, the set-theoretic explainable AI-based attribution score is presented in Section 5.3.4.

5.3.1 Architecture of 1D CNN

A conventional 1D CNN architecture primarily uses the convolution layer, pooling layer, and dense layer where the convolution and pooling layer extract the features and generate a reduced feature space. The dense layer transforms the reduced feature space from the previous layer to provide a combination of features for the classification of samples.

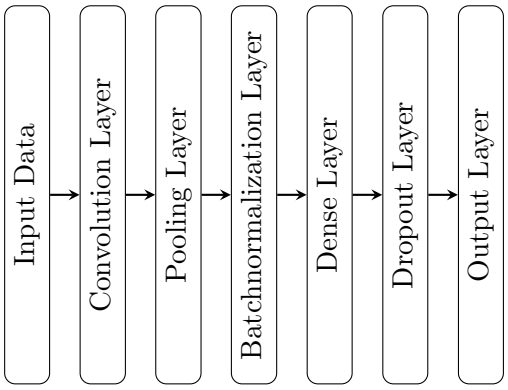


Figure 5.1: Schemetic Diagram representing the sequence of layers in MOHCNN.

A typical example of 1D CNN is shown in Fig. 5.2. To understand the working methodology, let us consider an input of 10×1 dimension and a filter of dimension 3×1 as an example. The convolution of the filter and input vector produces an 8×1 output by using a sliding window of 1. The product of the filter and input vector is provided to a non-linear activation function. The dimension of the output can be reduced by adding a pooling layer with a varying window size. In the example, the pooling size is considered as 2 which reduces the size of the output vector of the activation function to half. Two pooling operations are shown in Fig. 5.2, where the upper vector represents the output of max pooling and the lower one is the output of average pooling. The feature map is transformed into a one dimensional vector using a flattened layer. The output of the flatten layer is sent to the dense layer of activation functions which is connected to output layer. The output layer utilizes the 'softmax' function to classify patients.

Backpropagation concept is utilized in a conventional 1D CNN where the weights are changed recursively until the loss between the prediction and desired output is minimal. The loss is optimized by utilizing gradient descent techniques. Minimum the loss better the training of the CNN model. The output of the softmax function values, which are considered as the probabilities of the patients belonging to various classes, ranges from 0 to 1 and the sum of these values is 1. Certain modifications can result in improving the performance of the conventional 1D CNN such as changing the number of layers, finding suitable values for hyperparameters, choices of loss functions, requirement of regularization layers etc.

5.3.2 Architecture of MOHCNN

MOHCNN is developed to classify patients from various cancer classes using pan-cancer miRNA expression data. The expression profile of each patient for all miRNAs is fed directly as an input vector [131] to MOHCNN. The sequence of layers in the architecture of MOHCNN is shown in Fig. 5.1. While the convolutional and pooling layers are present

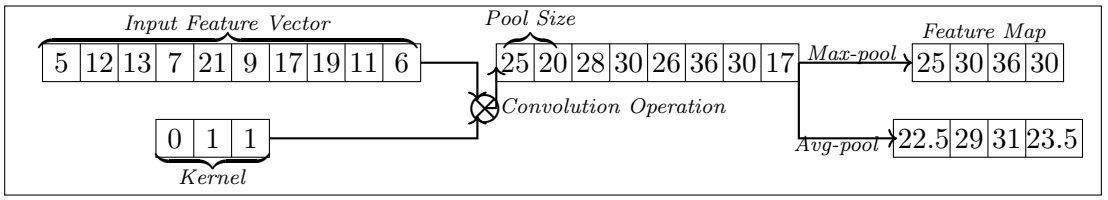


Figure 5.2: Convolutional and pooling operation of a 1D CNN using an input data of dimension 10.

as a part of conventional 1D CNN architecture, the batchnormalization and dropout layers are incorporated to increase the generalization and robustness of the developed model. The batchnormalization layer is used to normalize the distribution of neurons in the convolutional layer which helps in preventing the vanishing and exploding gradient problem in backpropagation [211]. While vanishing refers to the situation where gradients become too small and it prevents the network weights from changing, in exploding situation gradients become too large and it makes the model unstable. The dropout layer eliminates some neurons to avoid overfitting [212]. Note that, batchnormalization and dropout layers are used individually or in combination in [266–269], but in general, these are not used in conventional CNNs.

In MOHCNN, categorical cross-entropy loss function with $L2$ penalty is used as a loss function as the classification of patients using pan-cancer data is a multi-class problem. The loss function for the model is defined as:

$$L_{model} = CE(y^t, y^p) + \lambda \sum_{j=1}^{c} W_j^2 \tag{5.1}$$

where, $CE(y^t, y^p)$ represents the categorical cross entropy between true label (y^t) and predicted label (y^p), and $\lambda \sum_{j=1}^{c} W_j^2$ is the $L2$ penalty representing sum of regularization terms multiplied with user defined parameter λ. $CE(y^t, y^p)$ is further extended and the new loss equation is defined as:

$$L_{model} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{c} (y_{ij}^t * log(y_{ij}^p)) + \lambda \sum_{j=1}^{c} W_j^2 \tag{5.2}$$

where N and c represent the number of instances and classes, respectively.

5.3.3 Hyperparameter Optimization

In MOHCNN, the no of various types of layers, as well as other hyperparameters such as the number of filters, filter size, dropout rate, activation function, pooling size, and the number of units in the dense layer, are optimized in the multi-objective framework to improve the performance. The optimal set of solutions, based on non-dominated

solutions, and a trade-off between objectives is determined using the Pareto optimal front.

5.3.3.1 Objective Functions

In MOHCNN, the hyperparameter optimization problem is treated in a multi-objective framework using Bayesian optimization with a multivariate tree parzen estimator. The hyperparameter problem can be formulated as:

$$\begin{aligned} minimize\, f_m(x),\, m = 1, 2, 3; \\ f_1(x) = f_{train\_error} \\ f_2(x) = f_{validation\_error} \\ f_3(x) = f_{parameters} \end{aligned} \tag{5.3}$$

where the objective functions $f_1(x)$ and $f_2(x)$ are error values computed on training and validation data using the loss function defined as Eq. 5.2. The objective function $f_3(x)$ denotes the number of training parameters. All three objective functions are considered as minimization functions. The training and validation loss represent underfitting and overfitting of the model, two major concerns for using any deep learning model. The number of training parameters is minimized to reduce the computational complexity of the model.

5.3.3.2 Optimization

The hyperparameters of the developed architecture of MOHCNN are tuned using Bayesian optimization and CPCN data. Bayesian optimization tries to find the trials that determine the concurrence relations after a number of independent samplings. In hyperparameter optimization process, the number of filters, filter size, dropout rate, pooling size, and the number of units are varied from 16 to 256 in steps of 1, 2 to 5 in steps of 1, 0.1 to 0.7 in steps of 0.01, 2 to 5 in steps of 1, and 16 to 256 in steps of 1, respectively. The activation functions are chosen as sigmoid, relu, and leaky_relu, and each time, one of them is chosen randomly in the optimization process. The model is trained using 200 runs and each run consists of 100 trials. For each trial, 10 fold cross-validation is used, and in each fold, 50 epochs are used as inspired by existing studies [218,219]. Each trial provides a set of hyperparameter values, and hence, 200 runs provide $200 \times 100 = 20000$ set of solutions. A callback regularization technique is also used to stop the model if the errors do not change for five consecutive epochs.

Sixteen non-dominated sorted solutions, out of 20000 solutions, are obtained using the Pareto optimal front. Out of these 16 solutions, the solutions with minimum number of layers are first pointed out. Finally, a solution with minimum training error is chosen.

Table 5.2: Optimal hyperparameters for MOHCNN.

Layers	No. of Filters/Units	Filter Size	Activation Functions	Pooling Size	Dropout Rate
Convolution Layer1	18	5	Relu	4	-
Convolution Layer2	44	2	Relu	3	-
Dense Layer1	55	-	Leaky_relu	-	0.17
Dense Layer2	68	-	Leaky_relu	-	0.24

The chosen set of hyperparameters is shown in Table 5.2. It can be observed from the table that the number of convolutional layers and dense layers is 2. The number of filters and pooling size for convolutional layers are 18 & 44, and 2 & 3, respectively. The active function is Relu for convolution layers and leaky_relu for dense layers. The number of units in two dense layers is 55 and 68, and the dropout rate is 0.52 and 0.32. These hyperparameters are kept constant for other datasets as well.

5.3.4 Set-theoretic Explainable AI-based Attribution Score

A Set-theoretic explainable AI based attribution score (STEAAS) is developed to find the relevant miRNAs in a cancer class. It uses the concept of $Z-number$. The $Z-number$ is a two-element tuple where the first part represents the uncertainty of an element, and the second part denotes the reliability of the element for having that uncertainty [270]. In traditional $Z-number$ concept, probability is used as a reliability measure for determining the confidence in uncertainty. The $Z-number$ can be represented as:

$$Z-number = (uncertainty, reliability) \tag{5.4}$$

In STEAAS, the first tuple of $Z-number$ is called class score, and the second tuple is called reliability score. The class score of a miRNA is computed by considering the average of BayLIME (Bayesian Local Interpretable Model Agnostic Explanations) [149] and SHAP (Shapley Additive exPlainations) [150] values of the patients belonging to that class. BayLIME [149] is a modified version of LIME where Bayesian Ridge Regression is used as a regressor function instead of LASSO [133]. The BayLIME and SHAP values are then normalized within 0 to 1 for all the miRNAs. In reliability score, a class-dependent dispersion of each miRNA is computed using the GINI index which measures the inequality of expression profiles for all patients of a miRNA in a class. Finally, the class score and the reliability score of each miRNA for each cancer class are combined in a set-theoretic manner as an attribution score for selecting the relevant miRNAs. While, the details of class score involving BayLIME and SHAP are presented in Section 5.3.4.1 and 5.3.4.2, respectively, the details of reliability score are provided in Section 5.3.4.3.

5.3.4.1 Part of Class Score using BayLIME

LIME is a model-agnostic explanation technique that utilizes a linear surrogate method to explain a complex model locally [271]. The prediction of a classifier is used for training and explaining an instance by perturbing a dataset around it. The set of weight coefficients, β, generated using LIME can be defined as [271]:

$$\beta = \arg\min_{x \in G} L(f, g, \pi_p) + \omega(g) \tag{5.5}$$

where p denotes the instance to be explained, G is a set of g interpretable models, $L(f, g, \pi_p)$ indicates the local fidelity function which determines the accuracy in approximation provided by a particular model in the vicinity of π_p, f represents the deep learning model (in our case MOHCNN) to be explained, π_p is the kernel function which helps in assigning weights to the instances from the perturbed data, and $\omega(g)$ represents the interpretable model's complexity. The instances in the perturbed dataset around p are weighted based on a kernel function, and the patients having expression values closer to p are provided with higher weights. The BayLIME is a variant of LIME that uses the Bayesian principle for estimating the values of β [149].

In the case of pan-cancer expression data, patients are considered as instances and the miRNAs are considered as features which can be ranked based on their contributions in the predictions made by the MOHCNN. Some major steps of BayLIME to assign scores to the miRNAs for a particular patient are as follows:

S1) A patient p with m number of miRNAs is selected from the dataset for interpretation and it is considered as the original instance.

S2) A data perturbation around the patient p is performed and a new perturbed data $X' = x'_1, x'_2, \ldots, x'_n$, containing n patients, is generated. It can be represented as $X' = (x'_{ij}) \in \mathbb{R}^{n \times m}$ where i is the number of patients, j is the number of miRNAs, and $\mathbb{R}$ is a set of real numbers.

S3) The developed MOHCNN is trained with X' and class predictions by MOHCNN are noted as a prediction label vector $y' = [y'_1, y'_2, \ldots, y'_n]^T$.

S4) Weights to n patients of perturbed data X' are assigned according to their closeness to the patient p. The weights are calculated using the exponential kernel function and the new dataset is represented as (X'', y'').

S5) A linear regressor model is selected to train on the newly generated dataset, (X'', y''). The regressor model is represented as:

$$y'' = X''\beta + \epsilon \tag{5.6}$$

where β and ϵ represent the weight coefficients and Gaussian noise parameters, respectively.

S6) The steps from S1) to S5) are repeated to compute the β value of each patient in a particular cancer class.

S7) The average of β values, using all patients in a class, is then computed as:

$$B_j^c = \frac{\sum_{i=1}^{k} \beta_{ij}}{k} \tag{5.7}$$

where B_j^c is the score of jth miRNA in class c, β_{ij} is the score of jth miRNA corresponding to ith patient, and k denotes the number patients in a class. BayLIME utilizes the Bayesian ridge regression model (BRRM) instead of the ridge [272] and LASSO [133] models for regression. In BayLIME, the posterior estimates on β are a combination of the weighted sum of prior knowledge and new observations. This weighted sum may help in averaging out the randomness and improving the consistency of the model for the selection of miRNAs from the data.

5.3.4.2 Combining BayLIME and SHAP in Class Score

The concept of BayLIME is discussed in the Section 5.3.4.1 and SHapley Additive exPlanations (SHAP) is explained in Section 4.3.3 of Chapter 4. While SHAP is often used as both local and global explanation approach, BayLIME is only used for the former case. Further, both BayLIME and SHAP explanations are denoted as feature attribution methods. That view connects the BayLIME and SHAP to obtain interpretable features.

The contributions of a feature j (miRNA) in a class using BayLIME and SHAP are computed using 5.5 and Eqs. 4.9, respectively. The average of these values is considered as the class score for a miRNA. This is repeated for all the patients in the class. The class score of jth miRNA is defined as:

$$M_j = \left(\frac{\sum_{i=1}^{k} \beta_{ij} + \sum_{i=1}^{k} \chi_{ij}}{2k} \right) \tag{5.8}$$

where, β_{ij} and χ_{ij} denote the explanations of ith patient and jth miRNA obtained using BayLIME and SHAP, respectively, in a class. K denotes the number of patients in a cancer class.

In BayLIME and SHAP, different weighting schemes are used for the patients. While the patients are weighted based on their proximity to the selected patient in BayLIME, these are weighted based on the weights the 'coalition of miRNAs for a particular patient' obtains in the estimation of Shapley values. The lesser the number of 1's in the coalition vector, the smaller the weights in BayLIME. In SHAP, if the number of 1's in the coalition vector is few or many, then it achieves the highest weight [255]. In summary,

the class score represents the relevance of a miRNA in various cancer classes based on its role in the class prediction of patients during classification.

5.3.4.3 Reliability Score

The relevance of a miRNA in terms of class score is discussed in Section 5.3.4.2. However, the class score (See Eq. 5.8) does not represent the reliability of the miRNA for belonging to that particular class. To determine the reliability of a miRNA belonging to a class, probability can be computed using frequently occurring expressions of a miRNA in that class. Note that, in a practical situation, it is almost impossible to have the same expression values of a miRNA in two patients. In this scenario, the dispersion of expression values of a miRNA in a class is computed as a reliability score, which shows the relationship between different expression values. The dispersion shows the spread of expressions, and a lower spread indicates higher reliability of a miRNA to that class. The reliability score is computed using the Gini index.

Gini Index: In economics, the Gini index is utilized to determine the inequality of the distribution of wealth in a population [273]. Here, Gini index for a miRNA helps in finding the inequality of expression values in a class and it is defined as:

$$G_j = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|p_i - p_j|}{2n\sum_{i=1}^{n} p_i} \tag{5.9}$$

where p_i and p_j are the expression values of ith and jth patients for a miRNA. The Gini index is subtracted from 1 to compute the reliability score of a miRNA and it is defines as:

$$R_j = 1 - G_j \tag{5.10}$$

Some of the properties of the score are as follows: (i) $R_j \in [0, 1]$, where a value of 1 represents maximum reliability, and (ii) as the difference between the expression values decreases, the value of R_j increases.

In summary the set-theoretic explainable AI based attribution score of jth miRNA is represented as:

$$STEAAS_j = (M_j, R_j) \tag{5.11}$$

A miRNA with a high class score and a high reliability score is treated as a relevant miRNA in that cancer class. A threshold of 0.8 is chosen for both class score and reliability score. MiRNAs achieving both scores higher than the chosen thresholds are selected for validation using publicly available databases such as the ENCORI/Starbase and OncomiR.

The source codes of the MOHCNN-STEAAS and datasets related to the investigation are provided on Github page `https://github.com/joginder12/ICNNM`.

5.4 Experimental Results

In this section, we discuss the performance of MOHCNN. The hyperparameters are optimized as mentioned in Section 5.3.3.2 and then used to train the model. Seven datasets, CPCS, CPCNS, CNS, CPCCNS, breast, kidney, and lung, are used in the study. The results regarding the training and test performance of MOHCNN are reported in Section 5.4.1. The miRNAs selected by the developed STEAAS are biologically validated using the ENCORI/Starbase [153] and OncomiR [90] and reported in Section 5.4.2.

5.4.1 Performance of MOHCNN

The performance of MOHCNN, in terms of the number of training parameters and the training performance, is compared with some existing CNN models. Further, the comparison of MOHCNN with some existing CNN models, recent miRNA selection methods, and well established boosted classifiers is also provided in terms of test performance. The comparison methods are as follows:

- Three-layer CNN (ICNNM) [151], One layer CNN (Base-CNN) [131], stacked ensemble CNN (Ens-CNN) [132], and 6 layer CNN (SixL-CNN) [53]. These models are used for gene expression analysis but not for miRNA expression.
- Ensemble feature selection (EFS) [116], miRNA based pan-cancer diagnosis (SVM-RBF) [119], and an ensemble of 14 classifiers (Onco-Cls) [121]. These methods are used in previous studies for miRNA expression-based pan-cancer analysis.
- Gradient boosting (Gboost) [140] and Random Forest (RF) [140]. These two classifiers performed the best among various classifiers used for pan-cancer analysis in [116].
- Catboost (CBT) [148] and XGboost [152]. These are two highly cited state-of-the-art classifiers for tabular data.

Table 5.3: Comparing MOHCNN with related CNN models in terms of layers and parameters.

Measures	MOHCNN	Base-CNN	Ens-CNN	SixL-CNN
Number of convolution Layers	2	1	1	6
Number of parameters	33284	9,15,588	22,90,722	24, 65, 817

In Table 5.3, MOHCNN is compared with related CNN models in terms of training parameters. The minimum and maximum number of training parameters are 33284 and 2465817 for MOHCNN and SixL-CNN, respectively. Although the number of convolution layers is 2 for MOHCNN as compared to 1 for Base-CNN, the number of training

parameters for MOHCNN is approximately 30 times less than that of Base-CNN, which is the second method with the least training parameters.

Seven datasets CPCS, CPCNS, CPCCNS, CNS, breast, kidney, and lung, are used for comparing the classification performance of the MOHCNN. Each dataset is partitioned into training and test data where 70% of the data is used for training and 30% is kept for testing. The MOHCNN is trained using stratified 10-fold cross-validation. Nine folds are used for training and one fold for validation. After training, the model is tested on the remaining 30% of data to check the performance on unseen data.

The training performance of MOHCNN, for each fold, in terms of training accuracy, training error, validation accuracy, and validation error, and their averages for ten folds are reported in Table 5.4. The performance of a model is considered good if the accuracy is high and the error is low. The maximum average training and validation accuracies achieved by the MOHCNN are 1.00 for CNS, kidney and lung datasets and 0.99 for the same data, respectively. The minimum training and validation errors are 0.01 for the lung dataset and 0.02 for the kidney dataset, respectively.

Table 5.4: Training Performance of MOHCNN in terms of training accuracy (T_Acc), training error (T_Err), validation accuracy (Val_Acc), and validation error (Val_Err) for seven datasets.

Datasets	Measures	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	Average
CPCS	T_Acc	0.97	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	**0.99**
	T_Err	0.13	0.07	0.06	0.06	0.05	0.05	0.06	0.05	0.04	0.04	**0.06**
	Val_acc	0.91	0.95	0.95	0.98	0.99	0.98	0.98	0.97	0.98	0.99	**0.97**
	Val_err	0.35	0.18	0.18	0.10	0.07	0.07	0.07	0.10	0.10	0.03	**0.13**
CPCCNS	T_Acc	0.95	0.97	0.98	0.98	0.98	0.99	0.98	0.99	0.99	0.99	**0.98**
	T_Err	0.20	0.12	0.11	0.10	0.09	0.07	0.09	0.07	0.06	0.06	**0.10**
	Val_acc	0.89	0.90	0.96	0.95	0.97	0.98	0.99	0.98	0.98	0.99	**0.96**
	Val_err	0.46	0.43	0.17	0.22	0.14	0.09	0.07	0.12	0.12	0.05	**0.19**
CPCNS	T_Acc	0.96	0.97	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99	**0.98**
	T_Err	0.16	0.10	0.09	0.08	0.07	0.07	0.06	0.05	0.06	0.05	**0.08**
	Val_acc	0.91	0.95	0.96	0.97	0.98	0.98	0.98	0.99	0.99	0.99	**0.97**
	Val_err	0.37	0.18	0.16	0.11	0.09	0.08	0.07	0.06	0.05	0.04	**0.12**
CNS	T_Acc	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	**1.00**
	T_Err	0.14	0.05	0.03	0.02	0.02	0.01	0.02	0.01	0.01	0.02	**0.03**
	Val_acc	0.95	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	**0.99**
	Val_err	0.19	0.19	0.03	0.02	0.02	0.01	0.01	0.01	0.01	0.01	**0.05**
Breast	T_Acc	0.93	0.95	0.98	1.00	1.00	0.92	0.99	1.00	1.00	1.00	**0.98**
	T_Err	0.22	0.18	0.09	0.02	0.02	0.28	0.04	0.01	0.01	0.01	**0.09**
	Val_acc	0.63	0.91	0.94	0.99	0.94	0.84	0.95	1.00	1.00	1.00	**0.92**
	Val_err	1.24	0.28	0.18	0.04	0.12	0.38	0.13	0.02	0.01	0.01	**0.24**
Lung	T_Acc	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	**1.00**
	T_Err	0.04	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.00	**0.01**
	Val_acc	0.95	1.00	0.98	1.00	1.00	1.00	0.98	1.00	1.00	1.00	**0.99**
	Val_err	0.32	0.01	0.02	0.01	0.00	0.00	0.03	0.00	0.00	0.00	**0.04**
Kidney	T_Acc	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	**1.00**
	T_Err	0.03	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	**0.01**
	Val_acc	0.96	0.99	1.00	0.99	1.00	0.99	1.00	1.00	1.00	1.00	**0.99**
	Val_err	0.13	0.03	0.01	0.02	0.01	0.01	0.00	0.00	0.00	0.00	**0.02**

The comparison of MOHCNN with related CNN models in terms of the average of training accuracy, training error, validation accuracy, and validation error is reported in Table 5.5. The best results are marked in bold fonts. It is observed from the table that MOHCNN achieved the minimum average training and validation error and the

Table 5.5: Comparing the performance of MOHCNN with related CNN models in terms of training accuracy (T_Acc), training error (T_Err), validation accuracy (Val_Acc), and validation error (Val_Err) for seven datasets. The best results are marked in bold font.

Datasets	Methods	MOHCNN	Base-CNN	Ens-CNN	SixL-CNN
CPCS	T_Acc	**0.99**	0.92	0.97	0.95
	T_Err	**0.06**	0.31	0.09	0.18
	Val_Acc	**0.97**	0.90	0.95	0.88
	Val_Err	**0.13**	0.37	0.19	0.56
CPCNS	T_Acc	**0.98**	0.90	0.97	0.75
	T_Err	**0.08**	0.38	0.11	0.82
	Val_Acc	**0.97**	0.87	0.96	0.60
	Val_Err	**0.12**	0.47	0.16	3.95
CPCCNS	T_Acc	**0.98**	0.91	0.95	0.97
	T_Err	**0.10**	0.32	0.16	0.11
	Val_Acc	**0.96**	0.88	0.94	0.92
	Val_Err	**0.19**	0.47	0.22	0.40
CNS	T_Acc	**1.0**	0.85	0.99	0.95
	T_Err	**0.03**	1.18	0.05	0.19
	Val_Acc	**0.99**	0.59	0.98	0.78
	Val_Err	**0.05**	1.81	0.07	0.78
Breast	T_Acc	**0.98**	0.74	0.95	0.85
	T_Err	**0.09**	1.13	0.90	0.43
	Val_Acc	**0.92**	0.14	0.37	0.54
	Val_Err	**0.24**	1.90	0.34	1.42
Lung	T_Acc	**1.0**	0.92	0.99	0.56
	T_Err	**0.01**	1.04	0.12	1.38
	Val_Acc	**0.99**	0.76	**0.99**	0.46
	Val_Err	**0.04**	1.97	**0.04**	1.49
Kidney	T_Acc	**1.0**	0.91	0.97	0.74
	T_Err	**0.01**	0.84	0.09	0.75
	Val_Acc	**0.99**	0.85	0.97	0.61
	Val_Err	**0.02**	0.75	0.08	1.08

maximum average training and validation accuracy for all the datasets. For example, the training accuracy, training error, validation accuracy, and validation error values are 0.99, 0.06, 0.097, and 0.13, respectively, for MOHCNN using CPCS data, and the same values are obtained as 1.0, 0.01, 0.99, and 0.04 for lung data. Note that Ens-CNN also achieved the same validation accuracy (0.99) and error (0.04) for lung data. For other datasets, Ens-CNN achieved the second-best values for all four measures except the CPCCNS data, where SixL-CNN achieved the second-best training accuracy and error. Interestingly, MOHCNN achieved 100% training accuracy for CNS, lung, and kidney datasets.

The comparison of MOHCNN with the related methods and classifiers in terms of test accuracy, F-score, and MCC on 30% of test data is provided in Tables 5.6, 5.7, and 5.8, respectively. The best results achieved by any method for a dataset are represented in bold font.

It is observed from Table 5.6 that MOHCNN achieved the highest test accuracy values for all the datasets. The accuracy values are 1 for all the datasets except lung, where it is 0.99. Note that Base-CNN & Ens-CNN, Base-CNN & SixL-CNN, ICNNM, and SixL-CNN also achieved the highest test accuracy values for the CPCCNS, breast, CNS, and lung datasets, respectively. In general, Ens-CNN achieved the second-best results for all the datasets except the CPCS data.

Table 5.6: Comparing the performance of MOHCNN with related methods and boosted classifiers in terms of test accuracy for 30% of test data.

Datasets	Base-CNN	Ens-CNN	SixL-CNN	ICNNM	Onco-Cls	SVM-RBF	Gboost	XGboost	CBT	RF	MOHCNN
CPCS	0.97	0.96	0.90	0.99	0.95	0.92	0.89	0.93	0.94	0.91	**1.00**
CPNS	0.98	0.99	0.99	0.99	0.95	0.91	0.88	0.92	0.93	0.91	**1.00**
CPCCNS	**1.00**	**1.00**	0.98	0.99	0.94	0.88	0.86	0.92	0.93	0.90	**1.00**
CNS	0.97	0.97	0.96	**1.00**	0.96	0.77	0.89	0.93	0.94	0.91	**1.00**
Breast	**1.00**	0.96	**1.00**	0.98	0.83	0.75	0.78	0.79	0.80	0.76	**1.00**
Kidney	0.98	0.99	0.98	0.99	0.97	0.91	0.96	0.96	0.96	0.96	**1.00**
Lung	0.97	0.98	**0.99**	0.98	0.95	0.89	0.92	0.92	0.93	0.92	**0.99**

Table 5.7: Comparing the performance of MOHCNN with related methods and popular boosted classifiers in terms of F-score for 30% of test data.

Datasets	Base-CNN	Ens-CNN	SixL-CNN	ICNNM	Onco-Cls	SVM-RBF	Gboost	XGboost	CBT	RF	MOHCNN
CPCS	0.95	0.89	0.91	0.97	0.95	0.91	0.89	0.93	0.94	0.90	**0.97**
CPNS	0.94	0.91	0.93	**0.99**	0.95	0.89	0.88	0.92	0.93	0.90	0.97
CPCCNS	0.94	0.84	0.96	**0.98**	0.93	0.86	0.86	0.92	0.93	0.89	0.97
CNS	0.93	0.93	0.96	0.95	0.96	0.68	0.90	0.93	0.93	0.90	**0.98**
Breast	0.93	0.85	**1.00**	0.91	0.82	0.69	0.76	0.77	0.78	0.72	**1.00**
Kidney	0.95	0.94	0.94	0.97	0.93	0.89	0.96	0.96	0.96	0.96	**0.97**
Lung	0.95	0.96	0.95	0.98	0.94	0.88	0.92	0.92	0.93	0.92	**0.98**

It can be observed from Table 5.7 that MOHCNN achieved the highest F-score values except for the CPNS and CPCCNS datasets, where ICNNM outperforms MOHCNN. The values gained by ICNNM and MOHCNN for these datasets are 0.97 & 0.96 and 0.98 & 0.96 and it is noticed that the difference is insignificant. The values achieved by MOHCNN and ICNNM for CPCS, CPNS, CPCCNS, CNS, breast, kidney, and lung datasets are 0.97 & 0.97, 0.97 & 0.99, 0.97 & 0.98, 0.98 & 0.95, 1.0 & 0.91, 0.97 & 0.97, and 0.98 & 0.98, respectively.

Table 5.8: Comparing the performance of MOHCNN with related methods and popular boosted classifiers in terms of MCC for 30% of test data.

Datasets	Base-CNN	Ens-CNN	SixL-CNN	ICNNM	Onco-Cls	SVM-RBF	Gboost	XGboost	CBT	RF	MOHCNN
CPCS	0.94	0.89	0.87	0.96	0.93	0.89	0.89	0.88	0.93	0.91	**0.97**
CPNS	0.93	0.91	0.94	**0.97**	0.92	0.90	0.88	0.92	0.93	0.90	**0.96**
CPCCNS	0.95	0.85	**0.96**	**0.98**	0.94	0.88	0.85	0.92	0.93	0.90	**0.96**
CNS	0.93	0.92	0.94	0.94	0.93	0.75	0.88	0.93	0.93	0.90	**0.97**
Breast	**1.00**	0.83	**1.00**	0.90	0.73	0.61	0.64	0.66	0.69	0.61	**1.00**
Kidney	0.91	0.92	**0.96**	**0.97**	0.92	0.85	0.93	0.94	0.94	0.93	**0.96**
Lung	0.97	0.96	0.96	0.98	0.92	0.80	0.86	0.87	0.86	0.87	**0.98**

MOHCNN also obtained the best results for all the datasets except CPCNS, CPCCNS, and kidney datasets in terms of MCC value as provided in Table 5.8. The values for various datasets are 0.97, 0.96, 0.96, 0.97, 1.0, 0.96, and 0.98. For CPCNS, CPCCNS, and kidney datasets, ICNNM achieves higher results than MOHCNN, but the difference is insignificant. The third-best values are obtained by Base-CNN for most of the datasets except CNS and the kidney.

In summary, the average training and validation accuracy achieved by MOHCNN is 0.99 and 0.97 for all the datasets. The average training and validation accuracy achieved by Base-CNN, Ens-CNN, and SixL-CNN is 0.87 & 0.75, 0.97 & 0.96, and 0.82 & 0.69, respectively. The test performance of MOHCNN in terms of accuracy, F-score, and MCC varies from 0.99 to 1.00, 0.97 to 1.00, and 0.96 to 1.0, respectively. The MOHCNN performed better than the related methods in 215 out of 231 (3 measures $\times$ 7 datasets $\times$ 11 methods) cases.

5.4.2 Validation of miRNAs Selected using STEAAS

To determine the relevance of a miRNA in a cancer class, the developed class score and the reliability score in STEAAS are used. The miRNAs those obtained both class score and reliability score higher than the chosen threshold of 0.7 are selected as relevant ones. The class score and reliability score vary from 0 to 1, where a value of 1 for both scores indicates the maximum relevance of a miRNA in a class. The miRNAs selected for various cancer classes from CPCS data are reported in Table 5.9. Further, the miRNAs are also selected from subtype datasets and reported in Table 5.10.

Table 5.9: Biological validation of the selected miRNAs for CPCS dataset. The target genes for a miRNA corresponding to a particular cancer class are obtained using OncomiR and ENCORI/Starbase databases. The references mentioning the role of miRNAs in the corresponding cancer class are provided in the last column.

MiRNAS	Cancer class	Class Score	Reliability Score	Target Genes	References
hsa-mir-375	BLCA	1	0.81	GAS1, CALD1, QKI	[274]
	BRCA	0.88	0.91	BCORL1, SPINK13	
	ESCA	1	0.83	VPS13B	
	KICH	1	0.82	DCLK3, Z1C1	
	KIRP	1	0.84	SEP15	
	LUAD	0.81	0.93	GAS1, QKI, CALD1	
	OVAR	0.77	0.89	RASD1, FNDC3B, POU3F1	
	READ	0.93	0.95	GAS1,QKI	
	THYM	0.72	0.88	HTR5A	
	UCS	0.7	0.86	TET1, PYGO1	
	PAAD	1	0.93	NPAS3, CNOT7	
	PCPG	1	0.95	CAMSAP2	
	PRAD	1	0.98	HTR5A, XAF1	
hsa-mir-10b	CHOL	0.71	0.92	LANCL3, IL17RA	[275]
	LUAD	0.71	0.97	TM9SF3	[276]
	LUSC	0.71	0.98	C3ORF70	[276]
hsa-mir-192	READ	0.81	0.98	IGF2, BDNF	[274]
	STAD	0.75	0.95	CHML, ABL2, FAM162B	
hsa-mir-106a	ESCA	0.72	0.84	FAM49B, GTF2H1, PLEKHA8, PTPRF	[277]
hsa-mir-30a	LAML	0.8	0.89	FBX045, CCNF, CCNE2, MYBL2	[278]
hsa-mir-101-1	OVAR	0.71	0.89	LGI2, ADAMTS17, PHLDA1, ABHD17B	[279]
hsa-mir-153-2	READ	0.72	0.88	POLR2M, SH3BP4, SNAI1, RPS6KB1	[280]

In Table 5.9, the miRNAs selected by the developed STEAAS are reported for CPCS dataset. The target genes of the miRNAs, for the corresponding cancer classes, are

obtained using OncomiR [281] and ENCORI/Starbase [153] database. The selcetd miRNAs or their target genes are then validated with the help of existing biological studies. It is noticed from the first miRNA in the table that has-miR-375 achieved a class score greater than 0.70 in thirteen cancers. According to ENCORI database, has-miR-375 is differentially expressed in PRAD cancer with a p-value of $1.5e^{-}67$. It also achieved a class score of 1 and a reliability score of 0.95. The class score of the second miRNA, hsa-miR-10b, is 0.71 in CHOL, LUAD, and LUSC cancer classes. For the same classes, the reliability scores are 0.92, 0.97, and 0.98. The hsa-miR-10b is mentioned as an important marker in these cancers as per OncomiR database. It is also mentioned as a key biomarker in [275], [276], and [276] for CHOL, LUAD, and LUSC cancer, respectively. The third miRNA, hsa-miR-192, achieved a class score of 0.81 and a reliability scores of 0.98 for READ cancer class. The scores for the STAD cancer class are 0.75 and 0.95. This miRNa is mentioned as differentially expressed in both the cancer classes according to OncomiR and ENCORI databases. However, hsa-miR-192 is not mentioned as a key biomarker for these cancer classes in any existing literature. Therefore, the finding related to OncomiR and ENCORI databases suggests that hsa-miR-192 may be a novel biomarker for READ and STAD cancer class. The remaining four miRNAs, hsa-miR-106a, hsa-miR-30a, hsa-miR-101-1, and hsa-miR-153-2 are also mentioned as key biomarkers, for the mentioned cancer classes in Table 5.9, in OncomiR database as well as existing studies [277–280].

Table 5.10: Biological validation of the selected miRNAs for breast, lung and kidney datasets. The target genes for a miRNA corresponding to a particular cancer class are obtained using OncomiR and ENCORI/Starbase databases. The references mentioning the role of miRNAs in the corresponding cancer class are provided in the last column.

Dataset	MiRNA	Cancer	Class Score	Reliability Score	Targets genes	References
Lung	hsa-mir-455	LUSC	0.78	0.93	TMEFF1, CCNJ, PPFIA3, SLC17A6	[282]
	hsa-mir-375	LUSC	0.77	0.89	BCORL1, SPINK13	[154]
	hsa-mir-9	LUSC	0.75	0.89	PDCD4, ABCG1, MYRF, GSPT2,ARFRP1	[283]
	hsa-mir-125a	LUAD	0.78	0.96	FUT6, FBXL17, C12ORF5, JAK1, BCKDHA, FAM122A, MEGF11	[154]
	hsa-mir-505	LUAD	0.75	0.92	FBXL17, VGLL3, CHP1	[284]
	hsa-mir-652	LUAD	0.75	0.92	ATP11B	[285]
	hsa-mir-874	LUAD	0.77	0.89	USP8, KCNQ1, ZBTB7A, TEF, LPAR5	[286]
Kidney	hsa-mir-23a	KIRC	0.79	0.98	STAT5B, NDST1, MED12, CBLN1, MAP3K3	[287]
	hsa-mir-141	KICH	0.83	0.87	SDHC, UGT1A7, PNMA3, ALKBH5, AK2	[288]
	hsa-mir-625	KIRP	0.77	0.93	DNAJA1	[289]
Breast	hsa-mir-1307	HER2	0.75	0.96	RASSF2	[290]
	hsa-mir-130a	HER2	0.76	0.91	RASA1, SOCS6, ENPP5, ZFYVE26, ATG16L1, MECP2	[291]
	hsa-mir-140	HER2	0.76	0.97	KIAA0232, DSC1, CHD2, SRGAP3, SLC24A1, ZNF772, TBX18, ZNF740	[292]
	hsa-mir-144	HER2	0.78	0.87	TBX18, DST, CEP68, FAT4	[293]
	hsa-mir-146b	HER2	0.82	0.95	ANPEP, SEMA4C	[155]
	hsa-mir-30a	HER2	0.82	0.95	SCYL3, ZFYVE26, KIF16B, RASA1, ANKRA2	[155]
	hsa-mir-378c	HER2	0.81	0.94	JADE2, RBBP9, C2ORF88	[294]
	hsa-mir-191	LUMB	1.00	0.95	CREBBP	[295, 296]
	hsa-mir-3613	LUMA	0.93	0.88	C3ORF70, ATG16L1, DUSP18, KIAA0232	[297]

In Table 5.10, the miRNAs selected from breast, lung, and kidney datasets are reported along with their class score, reliability score, target genes, and references mentioning their role as biomarkers. In the lung dataset, three miRNAs, hsa-miR-455, hsa-miR-375, and hsa-miR-9, are selected for the LUSC subtype. Their class scores are 0.75, 0.77, and 0.75, and their reliability scores are 0.93, 0.89, and 0.89. These miRNAs are also found to be relevant as per the OncomiR database and existing investigations [154, 282, 283]. According to the OncomiR database, hsa-miR-125a is a key biomarker for the LUAD subtype with target genes FUT6, FBXL17, JAK1, etc. It is also found to be deregulated in LUAD as per [154]. Hsa-miR-505 and hsa-miR-652 are also mentioned as key markers in lung cancer [282, 284].

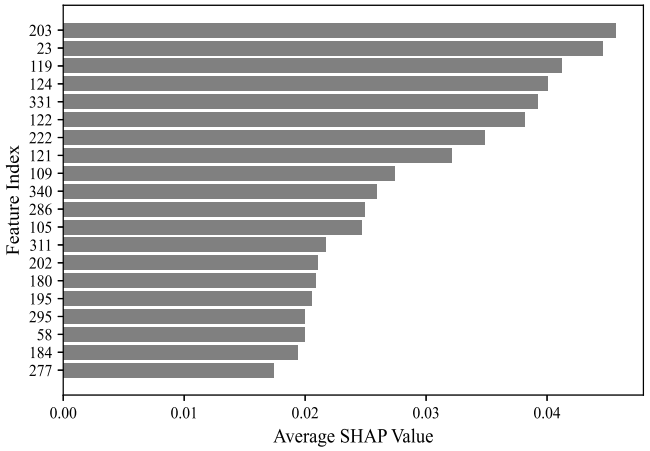
For the breast dataset, seven miRNAs are selected from the HER2 subtype, and one miRNA from each class LUMA and LUMB. The class score and the reliability score of the seven miRNAs from the HER2 subtype range from 0.75 to 0.82 and 0.87 to 0.97, respectively. These miRNAs are also validated using the Oncomir database and related literature, as mentioned in the table. Interestingly, hsa-miR-191 achieved a class score of 1 (which is the maximum score a miRNA can achieve) and a reliability score of 0.95 in the LUMB subtype. It is also reported as a biomarker in LUMB according to [296]. Similarly, hsa-miR-3613 is also mentioned as a biomarker in the LUMA subtype in [297].

For the kidney dataset, miRNAs hsa-mir-23a, hsa-mir-141, and hsa-mir-625 are selected for subtypes KIRC, KICH, and KIRP, respectively. Their class scores are 0.79, 0.83, and 0.77, and their reliability scores are 0.98, 0.87, and 0.93. These miRNAs are also mentioned as important biomarkers in the mentioned subtypes in various studies [287–289].

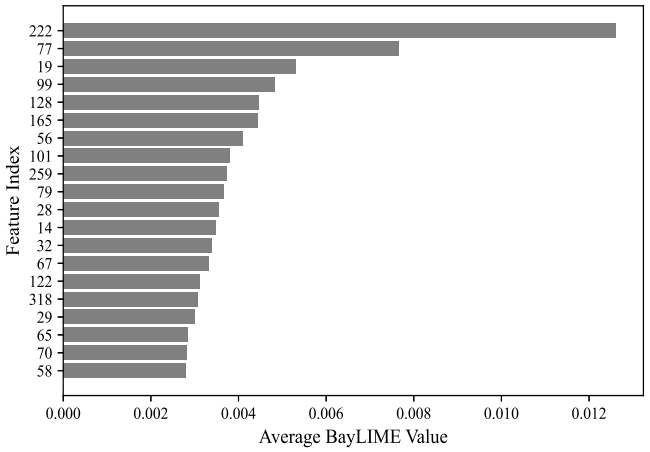
SHAP and BayLIME Visualizations: In Fig. 5.3, SHAP and BayLIME values for 10 patients from the ESCA class are plotted. The average SHAP and BayLIME values of the top 20 miRNAs are shown in Figs. 5.3(a) and 5.3(b), respectively. It is observed from the figure that the indices 222, 58, and 122 are common. The miRNA at index 222 is ranked first in BayLIME and seventh using SHAP scores. In the CPCS dataset, miR-375 is at 222 which is ranked top-most miRNA with a class score of 1 in the ESCA class. The miRNAs at 58 and 122 are 135b and 203b, respectively. These miRNAs are also selected as relevant ones using STEAAS with class scores of 0.68 and 0.68.

5.5 Complexity of MOHCNN-STEAAS

MOHCNN-STEAAS also utilizes 1D CNN model as a classifier for classifying patients from various cancer types. While the hyperparameters of the 1D CNN are optimized using Bayesian optimization in a single-objective framework in ICNNM, these are optimized in a multi-objective framework in MOHCNN. The discussion on the complexity of 1D CNN and SHAP is already provided in Chapter 4. Here, we provide a discussion



(a) Average SHAP values for top 20 features from ESCA class.



(b) Average BayLIME values for top 20 features from ESCA class.

Figure 5.3: The SHAP and BayLIME visualization of top ranked miRNAs.

on the complexity of the multiobjective tree-structured Parzen estimator for optimizing hyperparameters of a 1D CNN, Baylime to determine the scores of miRNAs based on the prediction of MOHCNN, Gini index to compute the reliability of the selected miRNA using class score, and then the total complexity of the developed model.

Multi-objective Optimization of hyperparameters using Bayesian Optimization with the Tree Parzen Structured Estimator: In the previous chapter, the time complexity of Bayesian optimization in single objective framework is provided. Here, we optimized the hyperparameters of a 1D CNN in a multi-objective framework. The complexity can be considered $O(H \times E \times o \times C_f)$, where o denotes the number of objective functions.

Class Score using SHAP and BayLIME: The time computation of SHAP is provided in Chapter 4. For BayLIME, it is mentioned in [149] that the complexity to

generate an explanation using an explainable model is dominated by the black box model which in this case is MOHCNN. Suppose a perturbed dataset of size P_D is generated to explain a patient with n miRNAs, MOHCNN will make predictions for all the samples in the perturbed dataset. This perturbation will be repeated for all the samples in the original dataset and the time complexity will be increased by a factor P_D to that of MOHCNN. Hence, the complexity of integrating BayLIME with MOHCNN in class score is $O(M \times P_D \times L \times f \times k \times n)$.

Reliability Score: The reliability score is computed using the Gini index, and the time required to compute the Gini Index for a miRNA is $O(M^2)$ as it computes the difference between each pair of values. For n miRNAs, the time complexity required for the reliability score, involving the Gini index, is $O(n \times M^2)$.

In summary, the computational complexity of the MOCNN and STEAAS after integration is $O(H \times E \times o) + O(M \times L \times f \times k \times n) + O(2^n) + O(M \times P_D \times L \times f \times k \times n) + O(n \times M^2) \approx O(H \times E \times o \times C_f) + O(M \times P_D \times C_f) + O(M \times S \times k \times C_f) + C_f$.

5.6 Discussion and Conclusion

In this study, a multi-objective framework for optimizing hyperparameters of a 1D CNN (MOHCNN) and a set-theoretic explainable AI-based attribution score (STEAAS) for identifying relevant miRNAs in various classes of cancer in pan-cancer data are developed. In the developed architecture, batch normalization after each pooling layer helps in preventing the vanishing and exploding gradient problem in backpropagation. Further, the addition of a dropout layer after each dense layer prevents the model from overfitting by stopping the neurons from becoming too specific for certain tasks. In the multi-objective paradigm, training error, validation error, and the number of training parameters are considered as objective functions. Bayesian optimization with multi variate tree parzen estimator is used for optimizing the hyperparameters using CPCNS data where the number of cancer classes is 33 and the number of normal classes is 16. As mentioned earlier, to avoid biases in the optimization of MOHCNN, 6 normal classes are removed from the original pan-cancer data (CPCN) to derive the CPCNS data as the number of normal patients in these classes is less than 5.

The performance of MOHCNN is evaluated on the CPCS, CPCNS, CNS, CPCCNS, breast, kidney, and lung datasets. The training and test results on different datasets with varying sample (patient) sizes show the stability and robustness of MOHCNN. It is observed from Tables 5.6, 5.7, and 5.8 that MOHCNN performs superior to the related methods, used for comparison, in 215 out of 231 cases. Further, MOHCNN requires only 33284 training parameters, which is approximately 30 times less than the Base-CNN, the second method with the least training parameters. The ranges of F-score and MCC values are 0.92 to 1.0, and 0.89 to 1,0, respectively, for MOHCNN. The average F-score

and MCC values are 0.97 and 0.96, respectively. Considering the results, the ICNNM is observed as the second-best method. MOHCNN performs better than ICNNM in 37 out of 42 cases with fewer training parameters. Hence, MOHCNN is more robust than the ICNNM.

In STEAAS, the class score and the reliability score help in identifying relevant miRNAs in various cancer classes. While the class score, computed using BayLIME and SHAP values, helps in selecting miRNAs based on their contribution towards the prediction of class labels of patients, the reliability score determines the inequality of expression values of a miRNA in a class. While ICNNM only uses SHAP to compute the attribution score of miRNAs, STEAAS utilizes the combination of the class score and the reliability score to select relevant miRNAs, which makes STEAAS more interpretable. Further, two explainable AI-based methods, BayLIME and SHAP, are integrated into the class score to improve the miRNA selection process. The miRNAs selected by STEAAS are validated using OncomiR and ENCORI/Starbase databases and related existing studies. Most of the miRNAs selected by STEAAS are mentioned as key biomarkers in various studies. For example, hsa-miR-125a and hsa-miR-30a are selected from lung and breast datasets and these are also mentioned as biomarkers in [154] and [155], respectively.

Chapter 6

Conclusion and Future Scope

Cancer can develop in any part of the body and can metastasize to other parts. Late diagnosis and anti-cancer drug resistance are major reasons for failure of cancer treatment and low survival rates of patients. A number of biochemical and biological investigations are conducted to establish the role of miRNAs in cancer treatment. Understanding the roles of miRNAs in cancer using wet lab based experiments is expensive. Although this problem can be handled computationally by overcoming certain challenges. The challenges in developing computational methods for identifying miRNAs in cancer are: i) not all miRNAs are important in the classification of cancer, and ii) a miRNA may express similarly in more than one cancer as discussed in Chapter 1 (See Section 1.4.1.3). To address these challenges, various computational methods are developed and discussed in four contributory chapters (Chapter 2-5) of this thesis. The problems handled in these chapters involve identifying important miRNAs for improving two-class (control and drug resistant patient) classification and multi-class cancer classification (pan-cancer analysis) problems. In chapter 1, a literature review and organization of the thesis are provided. In Chapter 2, two computational methods named EDWFC and HCEDFCR are developed for identifying drug resistant miRNAs using biological knowledge. In Chapter 3, an integration of fuzzy rough set based relevance and redundancy entropy measures using weights is introduced for identifying miRNAs that help in better classification of control and drug resistant patients. In Chapter 4, an interpretable convolutional neural network is developed for detecting miRNAs in various cancers using pan-cancer miRNA expression data. In Chapter 5, a 1D-CNN is optimized in multi-objective framework and a set theoretic explainable AI based attribution score is presented for identifying miRNAs in various cancers.

6.1 Conclusion

In Chapter 1, a review of the existing studies those deal with miRNAs associated with drug resistance in cancer and role of miRNAs in various cancers is provided. It is observed that the number of computational approaches is much lower than that of wet laboratory based studies. The lack of computational approaches for the identification of drug resistant miRNAs using expression data motivated us to develop three new methods (Chapter 2 and 3). Further, most of the existing computational methods in pan-cancer analysis are not optimized in performance and are unable to identify class specific miRNAs. These approaches focus on only classification of various cancers. Hence, two miRNA identification methods are developed in Chapters 4 and 5 for identifying miRNAs in various cancers in optimization framework.

Two methods (EDWFC and HCEDFCR) are developed for identifying drug resistant miRNAs in Chapter 2. In the EDWFC, the average rank of known drug resistant miRNAs are minimized by first multiplying the Euclidean distance and fold change between the averages of control and resistant expressions, and then varying the power of the fold change. A portion of miRNAs is selected from the top of the ranked list and used for classifying control and drug resistant patients. In the HCEDFCR, a histogram based clustering method is used to divide the miRNAs into different clusters. The product of the Euclidean distance and fold change value between the control and resistant miRNA expressions are used to rank the clusters and also miRNAs inside the clusters. MiRNAs in the top ranked cluster are considered as important. The key findings are:

- In EDWFC, Minimizing the rank of known miRNAs helps in finding an accurate power of fold change in the similarity measure to discriminate control and drug resistant expressions. This measure in turn helps in identifying unknown miRNAs with high discriminating capability and ranks them in the top order. Hence, the top miRNAs selected by the EDWFC are helpful in classifying control and drug resistant patients in a better manner. The comparison of EDWFC with some well-known techniques shows that the method performs the best in 302 out of 315 cases.

- In HCEDFCR, grouping of miRNAs is based on their frequency of expression magnitude (sum of expressions of patients) and then ranking those groups and miRNAs inside them. These help to detect top ranked drug resistant miRNAs from each group with miRNAs having similar overall magnitude of expression values across all drug resistant patients. For example, in colon FU & colon M datasets the known miRNA achieved the first position in the top cluster.

The EDWFC and HCEDFCR both utilize Euclidean distance and fold change for miRNA ranking. Further, the advantage of using biological knowledge about known

drug resistant miRNAs is demonstrated in EDWFC but not in HCEDFCR for miRNA ranking. In HCEDFCR, histogram based clustering is used in a unique way which uses sum of expression values for each miRNA.

In Chapter 3, two new z score based relevance and redundancy entropies are developed. These are integrated in a weighted framework, called WFIFRRRE, for selecting drug resistant miRNAs and classifying patients. The fuzzy membership of each miRNA expression value is computed using the ratio of z scores from one class to the other. These membership values are used to determine the fuzzy rough membership of expression values in lower and boundary regions. The average frequency of expression values of a miRNA in the aforesaid regions of fuzzy rough set is then determined to compute the relevance and redundancy entropies of each miRNA. The weight in WFIFRRRE, used to integrate relevance and redundancy entropies, is determined in a supervised manner where the F score is maximized in the classification performance in discriminating control and drug resistant patients for user defined number of selected miRNAs. The key findings are:

- Incorporating the weight in integrating the entropies and updating the weight to maximize the classification performance in terms of F score help in finding a set of important drug resistant miRNAs. Besides F score, the classification performance in terms of other performance measures such as sensitivity, specificity, accuracy, and MCC is also found to be better than related methods. In summary, WFIFRRRE performs better in 404 out of 416 cases while comparing it with related methods.
- Most of the miRNAs selected by WFIFRRRE from various datasets are validated using existing biological investigations. The remaining miRNAs can be validated in biological laboratories as the group of selected miRNAs achieved a high F-score in patient classification. For example, using the colon FU dataset, two out of six miRNAs (related F score 0.91) are found as key biomarkers in related literature. Hence, the remaining four miRNAs can be investigated in laboratories for their possible role in drug resistance.

The advantage of WFIFRRRE lies in combining two complementary components (fuzzy set and rough set) of soft computing. The judicious integration of these components not only helps in handling uncertainty in control and drug resistant class overlapping but also in determining the exactness of class size. Further, integration of relevance and redundancy entropies helps in finding a set of miRNAs that provides high classification accuracy. WFIFRRRE can also be applied for selecting other biomolecules whose expression data are available.

The developed methods are efficient for small sample size datasets. Moreover, these can also be utilized to handle the large datasets by partitioning the large dataset into

various small datasets and considering each combination of partitions. The developed scores can also be extended by integrating various distance measures that complement their properties, such as the multivariate mutual information measure, Spearman's correlation coefficient, shape shape-based coefficient methods.

In Chapter 4, an interpretable 1D convolution neural network model, called ICNNM, for pan-cancer analysis is developed. Some modifications in the conventional 1D CNN architecture are performed by adding a batchnormalization layer and a dropout layer in each of the convolution layers. The batchnormalization layer reduces the change in network parameters during training by setting the mean and variance of the input to a particular value. The dropout layer decreases the overfitting of the network while training is performed with small sample datasets, such as breast, lung, and kidney subclasses. Once the architecture of the CNN is defined, the hyperparameters are optimized using Bayesian optimization with a multivariate tree parzen estimator (BoMTPE). The optimized CNN as a classifier is then integrated with SHAP to compute attribution scores of miRNAs and rank them for each class of cancer in a single dataset. The key findings of ICNNM are as follows:

- The utilization of the dropout layer and batch normalization in the architecture improves the efficiency of the model in dealing with data of varying sample sizes from 817 to 11000 patients. The model performs better than the compared methods in 311 out of 315 cases.
- The attribution score of miRNAs computed using the expected gradient-based prior in terms of SHAP values identifies those miRNAs whose change in expression values resulted in the class prediction label. Further, the class discriminating power of the selected miRNAs is evaluated using UMAP projection, t-statistics, and web-based bioinformatics tools. Most of the selected miRNAs are found to be key biomarkers in various cancers as per existing literature and biological tools.

In summary, ICNNM uses two different approaches: an optimized 1D CNN model and SHAP value-based interpretation. While optimized 1D CNN efficiently classifies the patients in normal and various cancer classes, SHAP identifies the miRNAs through interpretation those are responsible for efficient classification. ICNNM serves the purpose of miRNA or gene selection from various cancers using pan-cancer expression data.

A multi-objective framework for optimizing hyperparameters of a 1D CNN (MOHCNN) and a set-theoretic explainable AI-based attribution scores (STEAAS) for miRNA selection are developed in Chapter 5. In MOHCNN, batch normalization after each pooling layer prevents the vanishing and exploding of gradients during the backpropagation process. Further, the addition of a dropout layer after each dense layer prevents the model from overfitting by stopping the neurons from becoming too specific for certain tasks. After fixing the architecture of the model, Bayesian optimization with a tree

Parzen estimator is utilized in a multi-objective framework. Three functions: training error, validation error, and number of training parameters, are used as objective functions for finding optimal hyperparameters. While the training and validation error controls the bias-variance trade-off of the model, the number of training parameters reduces the model's complexity. A set-theoretic explainable AI-based attribution score (STEAAS), which uses the $Z - number$ concept, is also developed to identify miRNAs from various cancer classes with reliability.

The developed MOHCNN is first utilized to classify patients from various classes, and then BayLIME and SHAP are integrated in STEAAS to find the class score of miRNAs. The class score helps in miRNA ranking in various cancer classes. The reliability score utilizes the Gini index to determine the inequality in the spread of expressions. The class score and the reliability score of each miRNA are combined in a set-theoretic manner using the $Z - number$ concept for the selection of miRNAs. The key findings of the developed method are as follows:

- The multi-objective optimization framework helps the model in achieving comparable or superior performance with fewer training parameters than the related methods. As discussed earlier, the model outperforms the related models with only 33,284 parameters. The results of the model are found to be superior in 215 out of 231 cases.
- The set-theoretic attribution score, defined using the $Z-number$ concept, identifies the miRNAs whose i) variation in expressions results in accurate class prediction of patients and ii) expression profiles are similar in a class. Further, the selected miRNAs are validated using web-based bioinformatics tools. Most of the selected miRNAs are found to be key biomarkers in various cancers.

In summary, the advantage of the developed framework lies in utilizing a multi-objective framework for the optimization of CNN and in developing an attribution score for identifying miRNAs in various cancer classes with reliability. While the class score computed using explainable models helps in selecting the miRNAs responsible for better patient classification, the reliability score determines the similarity of patients in a class. Hence, the developed framework can serve the purpose of detecting miRNAs, genes, or DNAs from various diseases using expression data.

6.2 Future Scope

The EDWFC and HCEDFCR use the product of the Euclidean distance and fold change, which are computed using the average of the expressions of miRNAs of the control and resistant classes. As mentioned in Chapter 2, this helps to explore the search space

using two different similarity measures based on the distance and ratio of expression values. Multiplying two linear functions results in a quadratic function, and applying weight as a power increases its nonlinearity beyond the quadratic relation, which helps in identifying non-linear relationships among data points. In the future, nonlinear distance measures like generalized K nearest neighbor [298] and gradient based large margin nearest neighbor (GB-LMNN) [299], and Spearman's correlation [300] to assesses monotonic relationships (whether linear or not) can be used to find similarities or distances between miRNA expressions. These measures can also be incorporated in HCEDFCR.

The entropies in WFIFRRRE are developed using the z-score and the Type I fuzzy set. It will be worthwhile to design the entropies using other scoring techniques, higher-order fuzzy sets such as Type II and Type III. Type II fuzzy set handles the uncertainty about the membership function, which is certain in Type I fuzzy sets. The membership function of a Type II fuzzy set uses one more dimension called "footprint of uncertainty". This additional dimension provides information about the uncertainty of a Type I fuzzy set in a better way [301]. Similarly, the Type III fuzzy set is more flexible and advantageous in handling uncertainty than both Type I and Type II. Further, the dependency of miRNAs selected by WFIFRRRE can be pointed out more accurately by incorporating he rating dependency criterion with relevance and redundancy. In general, features selected using fuzzy logic and rough set-based techniques are considered fairly interpretable. In Chapter 3, WFIFRRRE provides fuzzy logic and rough set-based scores to miRNAs which help to interpret and select them. On the other hand, deep learning-based methods are considered as black-box and hence explainable AI-based techniques are integrated with classifiers like CNNs in Chapters 4 and 5 to interpret and select them.. In future, WFIFRRRE can be extended to provide more interpretable features in two ways. One possible approach is by integrating WFIFRRRE with various fuzzy methods such as Technique for Order Preference by Similarity to Ideal Situation (TOPSIS) [302], VIse Kriterijumska Optimizacija I Kompromisno Rešenje (VIKOR) [303], hexagon of opposition [302], and graded hexagon of opposition [304]. In another way, the SVM classifier, in WFIFRRRE, can be enhanced by integrating explainable methods with it for selecting relevant miRNAs based on patient classification performance. Explainable SVM can be developed by integrating SHAP [255], LIME [255] or BayLIME [149] with SVM.

The methods, in Chapters 2 and 3, are efficient for both small sample size and large sample size datasets as these are developed using various classical techniques such as Euclidean distance, fold change, histogram, fuzzy logic and rough sets. The goal of the developed methods is to rank and select the miRNAs from the drug resistant dataset consisting of control and drug resistant classes. These methods can also be extended to select miRNAs from multiple classes. Two popular strategies for extension are *one-vs-rest* and *pairwise (one-vs-one)* decomposition approaches.

In *one-vs-rest* approach, the division is performed in such a way that each time

one class will be separated from all the other classes and the remaining classes will be combined into one class. For each separation, the methods developed in Chapters 2 and 3 can be applied to rank the miRNAs in each subclass. After completing the process for all the subclasses, the ranks of all the miRNAs can be aggregated. One vs the rest separation for pan-cancer data will result in highly imbalanced class data. This class imbalance problem handling will be challenging for the methods developed in Chapters 2 and 3. Hence, data augmentation process will be required. It is a technique used to generate artificial data to increase the size and diversity of the class with a lower number of samples. Data augmentation can be performed by rotating the data, repeating the data samples, adding noise to the data, etc.

In the *pairwise (one-vs-one)* decomposition approach, the multiclass problem is divided into $\frac{K(K-1)}{2}$ binary subclass problems, each consisting a pair of classes (C_i, C_j). The developed miRNA selection methods in Chapters 2 and 3 can be applied to each pair, generating a set of discriminative miRNAs that separate C_i from C_j. The miRNA relevance score can be computed by combining the pairwise scores, for instance, by averaging across all pairs where the miRNAs are selected or by weighting based on class prior probabilities.

In Chapter 4 and 5, interpretable CNN-based frameworks are utilized to classify patients from multiple cancer classes and to find the relevant miRNAs. In Chapter 4, a 1D CNN is optimized in terms of the number of layers as well as other hyperparameters in a single objective framework. The optimized model is applied to expression datasets in cancer. There are some other problems, like electroencephalogram (EEG) and electrocardiogram (ECG) signal-based patient classifications, which may be solved using a 1D CNN. It can be achieved by transfer learning (a learning process where a model that is developed and trained to do one task can be used again or changed to do a task that is similar but different), where the developed ICNNM and MOHCNN can be directly applied to EEG and ECG data without training from scratch. Optimization of 1D CNN can be studied in the context of these problems. Further, the developed SHAP value-based miRNA interpretation technique can also be applied to interpret the EEG and ECG signals.

In Chapter 5, a 1D CNN is optimized using a Bayesian technique in a multi-objective framework. Metaheuristic-based optimization techniques such as genetic algorithms [305], particle swarm [306], and whale optimization [307] can also be explored in this multi-objective framework. Moreover, the developed set-theoretic approach involves attribution scores and confidence levels, which are represented in numeric form. These can also be represented in linguistic forms such as preference or range of expressions.

In Chapters 4 and 5, the positions of layers are first fixed in CNN architecture, and then the optimization is performed on hyperparameters of the layers. In the future, the positions of layers, in addition to the values of hyperparameters, can be optimized.

This approach can be used in both single and multiobjective frameworks. Further, to interpret the miRNAs, SHAP is used in Chapter 4, and BayLIME, SHAP, and reliability score are utilized in Chapter 5. These techniques can be improved by finding appropriate priors and adding regularization and regressor models, which may help in stabilizing the feature attribution and in finding robust features. Apart from BayLIME and SHAP, some other interpretation methods, such as counterfactual explanations and individual conditional expectations, can also be investigated.

The methods in Chapters 4 and 5 are developed to handle pan-cancer data, where the maximum number of samples is 11,119. However, in Chapters 2 and 3, the maximum number of samples is 58 in the lymphoblastic datasets, which is considered a low sample size dataset. As mentioned earlier, CNNs and other deep learning frameworks require a huge amount of data for training and may suffer from overfitting when applied to small sample size datasets. To handle small sample size datasets, these methods may be modified by using regularization, data augmentation, and intra-class similarity kernels techniques, etc.

The methodologies discussed in this thesis focus on analyzing miRNA expression data. Other types of biological data, such as sequence data, histopathological imaging, and radiological imaging data, can also be studied individually or in an integrated way. The integration of one dataset with another provides complementary information to each other and helps in increasing the accuracy of results. Further, the individual datasets as well as the integrated data can be used for identifying the deregulated miRNAs at different stages of cancer [308] such as Stage 0 (means there's no cancer, only abnormal cells with the potential to become cancer), Stage I (means the cancer is small and only in one area), Stage II & III (cancer is larger and has grown into nearby tissues or lymph nodes) and Stage IV (cancer has spread to other parts of the body). It will be worthwhile to extend the developed methodologies in this thesis, such as integrating Euclidean distance with fold change, relevance and redundancy entropies, optimized CNNs, and attribution scores from SHAP and LIME, not only on integrated datasets but also for identifying miRNAs important for various stages of cancer.

Bibliography

[1] National Cancer Institute. Genomic data commons, TCGA Study Abbreviations, 2023. `https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations`, Last accessed on 2023-11-30.

[2] Rebecca L Siegel, Angela N Giaquinto, and Ahmedin Jemal. Cancer statistics, 2024. *CA: a cancer journal for clinicians*, 74(1), 2024.

[3] Fengju Chen, Darshan S Chandrashekar, Sooryanarayana Varambally, and Chad J Creighton. Pan-cancer molecular subtypes revealed by mass-spectrometry-based proteomic characterization of more than 500 human cancers. *Nature communications*, 10(1):5679, 2019.

[4] Adam Abeshouse et al. The molecular taxonomy of primary prostate cancer. *Cell*, 163(4):1011–1025, 2015.

[5] Armando E Giuliano et al. Breast cancer—major changes in the american joint committee on cancer eighth edition cancer staging manual. *CA: a cancer journal for clinicians*, 67(4):290–303, 2017.

[6] George A Calin and Carlo M Croce. Microrna signatures in human cancers. *Nature reviews cancer*, 6(11):857–866, 2006.

[7] Indian Council of Medical Research-National Centre for Disease Informatics and Research (ICMR-NCDIR). World cancer day 2024: Close the care gap- addressing cancer care in india, 2024. `https://ncdirindia.org/display/wcd.aspx`, Last accessed on 2024-5-30.

[8] Giovanni Ciriello et al. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, 163(2):506–519, 2015.

[9] George A Calin and Carlo M Croce. Microrna signatures in human cancers. *Nature reviews cancer*, 6(11):857–866, 2006.

[10] Colin C Pritchard, Heather H Cheng, and Muneesh Tewari. Microrna profiling: approaches and considerations. *Nature Reviews Genetics*, 13(5):358–369, 2012.

[11] Carlo M Croce. Causes and consequences of microrna dysregulation in cancer. *Nature reviews genetics*, 10(10):704–714, 2009.

[12] Yuan Li, Zhenning Wang, Jaffer A Ajani, and Shumei Song. Drug resistance and cancer stem cells. *Cell Communication and Signaling*, 19(1):1–11, 2021.

[13] Junfang Ji et al. Microrna expression, survival, and response to interferon in liver cancer. *New England Journal of Medicine*, 361(15):1437–1447, 2009.

[14] George Adrian Calin et al. Frequent deletions and down-regulation of micro-rna genes mir15 and mir16 at 13q14 in chronic lymphocytic leukemia. *Proceedings of the national academy of sciences*, 99(24):15524–15529, 2002.

[15] Patrick S Mitchell et al. Circulating microrna as stable blood-based markers for cancer detection. *Proceedings of the National Academy of Sciences*, 105(30):10513–10518, 2008.

[16] Andrea Feliciano et al. Five microrna in serum are able to differentiate breast cancer patients from healthy individuals. *Frontiers in Oncology*, 10:586268, 2020.

[17] Yoontae Lee, Minju Kim, Jinju Han, Kyu-Hyun Yeom, Sanghyuk Lee, Sung Hee Baek, and V Narry Kim. Microrna genes are transcribed by rna polymerase ii. *The EMBO journal*, 23(20):4051–4060, 2004.

[18] Rui Yi, Yi Qin, Ian G Macara, and Bryan R Cullen. Exportin-5 mediates the nuclear export of pre-microrna and short hairpin rna. *Genes & development*, 17(24):3011–3016, 2003.

[19] miRBase. mirbase database, 2023. `https://mirbase.org/`, Last accessed on 2025-07-28.

[20] Nannan Zhang, Guowu Hu, Timothy G Myers, and Peter R Williamson. Protocols for the analysis of microrna expression, biogenesis, and function in immune cells. *Current protocols in immunology*, 126(1):e78, 2019.

[21] Chad J Creighton, Jeffrey G Reid, and Preethi H Gunaratne. Expression profiling of microrna by deep sequencing. *Briefings in bioinformatics*, 10(5):490–497, 2009.

[22] Eric A Miska et al. Microarray analysis of microrna expression in the developing mammalian brain. *Genome biology*, 5:1–13, 2004.

[23] Ralf Herwig et al. Large-scale clustering of cdna-fingerprinting data. *Genome research*, 9(11):1093–1105, 1999.

[24] J MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*, 1967.

[25] Sanghamitra Bandyopadhyay and Sankar Kumar Pal. *Classification and learning using genetic algorithms: applications in bioinformatics and web intelligence*. Springer Science & Business Media, 2007.

[26] Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[27] Teuvo Kohonen. Essentials of the self-organizing map. *Neural networks*, 37:52–65, 2013.

[28] Pablo Tamayo et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences*, 96(6):2907–2912, 1999.

[29] Ulrike Von Luxburg. A tutorial on spectral clustering, 2007.

[30] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

[31] Desmond J Higham, Gabriela Kalna, and Milla Kibble. Spectral clustering and its use in bioinformatics. *Journal of computational and applied mathematics*, 204(1):25–37, 2007.

[32] Emrah Hancer, Bing Xue, and Mengjie Zhang. Differential evolution for filter feature selection based on information theory and feature ranking. *Knowledge-Based Systems*, 140:103–119, 2018.

[33] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.

[34] Petra Leidinger et al. High-throughput mirna profiling of human melanoma blood samples. *BMC cancer*, 10:1–11, 2010.

[35] Jayanta Kumar Pal, Shubhra Sankar Ray, and Sankar K Pal. Identifying relevant group of mirnas in cancer using fuzzy mutual information. *Medical & biological engineering & computing*, 54(4):701–710, 2016.

[36] Pradipta Maji and Sankar K Pal. Fuzzy–rough sets for information measures and selection of relevant genes from microarray data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(3):741–752, 2009.

[37] Jayanta Kumar Pal, Shubhra Sankar Ray, and Sankar K Pal. Identifying drug resistant mirnas using entropy based ranking. *IEEE/ACM transactions on computational biology and bioinformatics*, 18(3):973–984, 2019.

[38] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422, 2002.

[39] Pavel Pudil, Jana Novovičová, and Josef Kittler. Floating search methods in feature selection. *Pattern recognition letters*, 15(11):1119–1125, 1994.

[40] Shib Sankar Bhowmick and Debotosh Bhattacharjee. Microrna-based cancer classification using feature selection wrapper. In *Advanced Computing and Systems for Security: Volume 14*, pages 197–209. Springer, 2021.

[41] Lin Sun, Xianglin Kong, Jiucheng Xu, Zhan'ao Xue, Ruibing Zhai, and Shiguang Zhang. A hybrid gene selection method based on relieff and ant colony optimization algorithm for tumor classification. *Scientific reports*, 9(1):8978, 2019.

[42] Huijuan Lu, Junying Chen, Ke Yan, Qun Jin, Yu Xue, and Zhigang Gao. A hybrid feature selection algorithm for gene expression data classification. *Neurocomputing*, 256:56–62, 2017.

[43] Jayanta Kumar Pal, Shubhra Sankar Ray, and Sankar K Pal. Fuzzy mutual information based grouping and new fitness function for pso in selection of mirnas in cancer. *Computers in biology and medicine*, 89:540–548, 2017.

[44] Li-Yeh Chuang, Cheng-Huei Yang, Kuo-Chuan Wu, and Cheng-Hong Yang. A hybrid feature selection method for dna microarray data. *Computers in biology and medicine*, 41(4):228–237, 2011.

[45] Lokeswari Venkataramana, Shomona Gracia Jacob, Rajavel Ramadoss, Dodda Saisuma, Dommaraju Haritha, and Kunthipuram Manoja. Improving classification accuracy of cancer types using parallel hybrid feature selection on microarray gene expression data. *Genes & genomics*, 41:1301–1313, 2019.

[46] Richard G Brereton and Gavin R Lloyd. Support vector machines for classification and regression. *Analyst*, 135(2):230–267, 2010.

[47] Carl Gold and Peter Sollich. Model selection for support vector machine classification. *Neurocomputing*, 55(1-2):221–249, 2003.

[48] Jair Cervantes et al. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 408:189–215, 2020.

[49] Thomas Bayes. Naive bayes classifier. *Article Sources and Contributors*, pages 1–9, 1968.

[50] Fabian Pedregosa et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[51] Mahesh Pal. Random forest classifier for remote sensing classification. *International journal of remote sensing*, 26(1):217–222, 2005.

[52] Michael W Browne. Cross-validation methods. *Journal of mathematical psychology*, 44(1):108–132, 2000.

[53] Mohanad Mohammed et al. A stacking ensemble deep learning approach to cancer type classification based on tcga data. *Scientific reports*, 11(1):15626, 2021.

[54] Shiqin Liu. Leave-p-out cross-validation test for uncertain verhulst-pearl model with imprecise observations. *IEEE Access*, 7:131705–131709, 2019.

[55] Alain Celisse and Stéphane Robin. Nonparametric density estimation by exact leave-p-out cross-validation. *Computational Statistics & Data Analysis*, 52(5):2350–2368, 2008.

[56] Atesh Koul, Cristina Becchio, and Andrea Cavallo. Cross-validation approaches for replicability in psychology. *Frontiers in psychology*, 9:1117, 2018.

[57] Sashikanta Prusty, Srikanta Patnaik, and Sujit Kumar Dash. Skcv: Stratified k-fold cross-validation on ml classifiers for predicting cervical cancer. *Frontiers in Nanotechnology*, 4:972421, 2022.

[58] Hualong Yu, Changyin Sun, Wankou Yang, Sen Xu, and Yuanyuan Dan. A review of class imbalance learning methods in bioinformatics. *Current Bioinformatics*, 10(4):360–369, 2015.

[59] Fuxiao Xin et al. Computational analysis of microrna profiles and their target genes suggests significant involvement in breast cancer antiestrogen resistance. *Bioinformatics*, 25(4):430–434, 2009.

[60] Haixiu Yang, Yanjun Xu, Desi Shang, Hongbo Shi, Chunlong Zhang, Qun Dong, Yizheng Zhang, Ziyi Bai, Shujun Cheng, and Xia Li. ncdrmarker: a computational method for identifying non-coding rna signatures of drug resistance based on heterogeneous network. *Annals of Translational Medicine*, 8(21):1–31, 2020.

[61] Fatemeh Ahmadi Moughari and Changiz Eslahchi. A computational method for drug sensitivity prediction of cancer cell lines based on various molecular information. *PloS one*, 16(4):1–31, 2021.

[62] Beata Smolarz, Adam Durczyński, Hanna Romanowicz, Krzysztof Szyłło, and Piotr Hogendorf. mirnas in cancer (review of literature). *International journal of molecular sciences*, 23(5):2805, 2022.

[63] Jun Lu et al. Microrna expression profiles classify human cancers. *nature*, 435(7043):834–838, 2005.

[64] Stefano Volinia et al. A microrna expression signature of human solid tumors defines cancer gene targets. *Proceedings of the National Academy of Sciences*, 103(7):2257–2261, 2006.

[65] Rani Faryal. Role of mirnas in breast cancer. *Asian Pac J Cancer Prev*, 12:3175–3180, 2011.

[66] Yoshimasa Saito, Hidekazu Suzuki, and Toshifumi Hibi. The role of micrornas in gastrointestinal cancers. *Journal of gastroenterology*, 44:18–22, 2009.

[67] Rahul Nagadia and other. mirnas in head and neck cancer revisited. *Cellular Oncology*, 36:1–7, 2013.

[68] Xin Xu et al. The role of micrornas in hepatocellular carcinoma. *Journal of Cancer*, 9(19):3557, 2018.

[69] David Petillo et al. Microrna profiling of human kidney cancer subtypes. *International journal of oncology*, 35(1):109–114, 2009.

[70] Tyler E Miller, Kalpana Ghoshal, Bhuvaneswari Ramaswamy, Satavisha Roy, Jharna Datta, Charles L Shapiro, Samson Jacob, and Sarmila Majumder. Microrna-221/222 confers tamoxifen resistance in breast cancer by targeting p27kip1*. *Journal of biological chemistry*, 283(44):29897–29903, 2008.

[71] Pichiorri et al. In vivo ncl targeting affects breast cancer aggressiveness through mirna regulation. *Journal of Experimental Medicine*, 210(5):951–968, 2013.

[72] Núria Mencia, Elisabet Selga, Véronique Noé, and Carlos J Ciudad. Underexpression of mir-224 in methotrexate resistant human colon cancer cells. *Biochemical pharmacology*, 82(11):1572–1582, 2011.

[73] Ken Kurokawa et al. Role of mir-19b and its target mrnas in 5-fluorouracil resistance in colon cancer cells. *Journal of gastroenterology*, 47(8):883–895, 2012.

[74] Smriti Kumar, Arooshi Kumar, Parag P Shah, Shesh N Rai, Siva K Panguluri, and Sham S Kakar. Microrna signature of cis-platin resistant vs. cis-platin sensitive ovarian cancer cell lines. *Journal of ovarian research*, 4(1):1–11, 2011.

[75] Kazuhiro Kitamura, Masahiro Seike, Tetsuya Okano, Kuniko Matsuda, Akihiko Miyanaga, Hideaki Mizutani, Rintaro Noro, Yuji Minegishi, Kaoru Kubota, and Akihiko Gemma. Mir-134/487b/655 cluster regulates tgf-b–induced epithelial–mesenchymal transition and drug resistance to gefitinib by targeting magi2 in lung adenocarcinoma cells. *Molecular Cancer Therapeutics*, 13(2):444–453, 2014.

[76] Richard Hummel et al. Microrna signatures in chemotherapy resistant esophageal cancer cell lines. *World journal of gastroenterology: WJG*, 20(40):14904, 2014.

[77] Diana Schotte, Renée X De Menezes, Farhad Akbari Moqadam, Leila Mohammadi Khankahdani, Ellen Lange-Turenhout, Caifu Chen, Rob Pieters, and Monique L Den Boer. Microrna characterize genetic diversity and drug resistance in pediatric acute lymphoblastic leukemia. *haematologica*, 96(5):703, 2011.

[78] Richard Ottman, Camha Nguyen, Robert Lorch, and Ratna Chakrabarti. Microrna expressions associated with progression of prostate cancer cells to antiandrogen therapy resistance. *Molecular cancer*, 13(1):1–21, 2014.

[79] Chang Hee Kim, Hark K Kim, R Luke Rettig, Joseph Kim, Eunbyul T Lee, Olga Aprelikova, Il J Choi, David J Munroe, and Jeffrey E Green. mirna signature associated with outcome of gastric cancer patients following chemotherapy. *BMC medical genomics*, 4(1):1–14, 2011.

[80] Wanzhong Yin, Ping Wang, Xin Wang, Wenzhi Song, Xiangyan Cui, Hong Yu, and Wei Zhu. Identification of micrornas and mrnas associated with multidrug resistance of human laryngeal cancer hep-2 cells. *Brazilian Journal of Medical and Biological Research*, 46:546–554, 2013.

[81] Palagani et al. Ectopic microrna-150-5p transcription sensitizes glucocorticoid therapy response in mm1s multiple myeloma cells but fails to overcome hormone therapy resistance in mm1r cells. *PLoS One*, 9(12):1–30, 2014.

[82] J Ma, C Dong, and C Ji. Microrna and drug resistance. *Cancer gene therapy*, 17(8):523–531, 2010.

[83] Mehri Ghasabi, Behzad Mansoori, Ali Mohammadi, Pascal HG Duijf, Navid Shomali, Naghmeh Shirafkan, Ahad Mokhtarzadeh, and Behzad Baradaran. Micrornas in cancer drug resistance: Basic evidence and clinical applications. *Journal of cellular physiology*, 234(3):2152–2168, 2019.

[84] Wei Zhang and M Eileen Dolan. Emerging role of micrornas in drug response. *Current opinion in molecular therapeutics*, 12(6):695–702, 2010.

[85] Umar Raza, Jitao David Zhang, and Özgür Şahin. Micrornas: master regulators of drug resistance, stemness, and metastasis. *Journal of molecular medicine*, 92(4):321–336, 2014.

[86] Wengong Si, Jiaying Shen, Huilin Zheng, and Weimin Fan. The role and mechanisms of action of micrornas in cancer drug resistance. *Clinical epigenetics*, 11(1):1–24, 2019.

[87] Mario Deng, Johannes Brägelmann, Ivan Kryukov, Nuno Saraiva-Agostinho, and Sven Perner. Firebrowser: an r client to the broad institute's firehose pipeline. *Database*, 2017:baw160, 2017.

[88] University of California. Santa cruz, GDC Pan-Cancer (PANCAN) Data, 2023. https://xenabrowser.net/datapages/, Last accessed on May-2023.

[89] Yu Hu, Hayley Dingerdissen, Samir Gupta, Robel Kahsay, Vijay Shanker, Quan Wan, Cheng Yan, and Raja Mazumder. Identification of key differentially expressed micrornas in cancer patients through pan-cancer analysis. *Computers in biology and medicine*, 103:183–197, 2018.

[90] Nathan W Wong, Yuhao Chen, Shuai Chen, and Xiaowei Wang. Oncomir: an online resource for exploring pan-cancer microrna dysregulation. *Bioinformatics*, 34(4):713–715, 2018.

[91] Olivia D Lara et al. Pan-cancer clinical and molecular analysis of racial disparities. *Cancer*, 126(4):800–807, 2020.

[92] Amir Sabbaghian et al. A panel of blood-derived mirnas with a stable expression pattern as a potential pan-cancer detection signature. *Frontiers in Molecular Biosciences*, 9:1030749, 2022.

[93] Peng Wu, Chaoqi Zhang, Xiaoya Tang, Dongyu Li, Guochao Zhang, Xiaohui Zi, Jingjing Liu, Enzhi Yin, Jiapeng Zhao, Pan Wang, et al. Pan-cancer characterization of cell-free immune-related mirna identified as a robust biomarker for cancer diagnosis. *Molecular Cancer*, 23(1):31, 2024.

[94] Jayanta Kumar Pal, Shubhra Sankar Ray, and Sankar K Pal. Identifying relevant group of mirnas in cancer using fuzzy mutual information. *Medical & biological engineering & computing*, 54(4):701–710, 2016.

[95] Roy Navon, Hui Wang, Israel Steinfeld, Anya Tsalenko, Amir Ben-Dor, and Zohar Yakhini. Novel rank-based statistical methods reveal micrornas with differential expression in multiple cancer types. *PloS one*, 4(11):1–10, 2009.

[96] Yongjun Piao, Minghao Piao, and Keun Ho Ryu. Multiclass cancer classification using a feature subset-based ensemble from microrna expression profiles. *Computers in biology and medicine*, 80:39–44, 2017.

[97] Srinivasulu Yerukala Sathipati and Shinn-Ying Ho. Identifying a mirna signature for predicting the stage of breast cancer. *Scientific reports*, 8(1):1–11, 2018.

[98] Piyushkumar A. Mundra and Jagath C. Rajapakse. Svm-rfe with mrmr filter for gene selection. *IEEE Transactions on NanoBioscience*, 9(1):31–37, 2010.

[99] Alok Sharma, Seiya Imoto, Satoru Miyano, and Vandana Sharma. Null space based feature selection method for gene expression data. *International Journal of Machine Learning and Cybernetics*, 3(4):269–276, 2012.

[100] Mohamed F Ghalwash, Xi Hang Cao, Ivan Stojkovic, and Zoran Obradovic. Structured feature selection using coordinate descent optimization. *BMC bioinformatics*, 17(1):1–14, 2016.

[101] Shun Guo, Donghui Guo, Lifei Chen, and Qingshan Jiang. A centroid-based gene selection method for microarray data classification. *Journal of theoretical biology*, 400:32–41, 2016.

[102] Saeid Azadifar et al. Graph-based relevancy-redundancy gene selection method for cancer diagnosis. *Computers in Biology and Medicine*, 147:105766–105779, 2022.

[103] Sukriti Roy, Joginder Singh, and Shubhra Sankar Ray. Weighted combination of łukasiewicz implication and fuzzy jaccard similarity in hybrid ensemble framework (wclfjhef) for gene selection. *Computers in Biology and Medicine*, 170:107981, 2024.

[104] Mahmoo Khalsan et al. A novel fuzzy classifier model for cancer classification using gene expression data. *IEEE Access*, 2023.

[105] Ansuman Kumar and Anindya Halder. Ensemble-based active learning using fuzzy-rough approach for cancer sample classification. *Engineering Applications of Artificial Intelligence*, 91:103591, 2020.

[106] Meng Hu et al. Attribute reduction based on neighborhood constrained fuzzy rough sets. *Knowledge-Based Systems*, 274:110632, 2023.

[107] Yanqing Niu et al. Mirna-drug resistance association prediction through the attentive multimodal graph convolutional network. *Frontiers in pharmacology*, 12:799108, 2022.

[108] Tongsen Zheng, Jiabei Wang, Xi Chen, and Lianxin Liu. Role of microrna in anticancer drug resistance. *International journal of cancer*, 126(1):2–10, 2010.

[109] Kai Zheng et al. Nasmdr: a framework for mirna-drug resistance prediction using efficient neural architecture search and graph isomorphism networks. *Briefings in Bioinformatics*, 23(5):bbac338, 2022.

[110] Jinhang Wei et al. Gcfmcl: predicting mirna-drug sensitivity using graph collaborative filtering and multi-view contrastive learning. *Briefings in Bioinformatics*, 24(4):bbad247, 2023.

[111] Yong-Jian Guan et al. Mfidma: a multiple information integration model for the prediction of drug–mirna associations. *Biology*, 12(1):41, 2022.

[112] Fanrong Yu et al. Psrr: a web server for predicting the regulation of mirnas expression by small molecules. *Frontiers in Molecular Biosciences*, 9:817294, 2022.

[113] Yan Zhao et al. Snmfsmma: using symmetric nonnegative matrix factorization and kronecker regularized least squares to predict potential small molecule-microrna association. *RNA biology*, 17(2):281–291, 2020.

[114] Junliang Liu et al. Hggn: Prediction of microrna-mediated drug sensitivity based on interpretable heterogeneous graph global-attention network. *Future Generation Computer Systems*, 160:274–282, 2024.

[115] Eti Meiri et al. A second-generation microrna-based assay for diagnosing tumor tissue origin. *The oncologist*, 17(6):801–812, 2012.

[116] Alejandro Lopez-Rincon et al. Automatic discovery of 100-mirna signature for cancer classification using ensemble feature selection. *BMC bioinformatics*, 20:1–17, 2019.

[117] Srinivasulu Yerukala Sathipati, Ming-Ju Tsai, Sanjay K Shukla, and Shinn-Ying Ho. Artificial intelligence-driven pan-cancer analysis reveals mirna signatures for cancer stage prediction. *Human Genetics and Genomics Advances*, 4(3), 2023.

[118] Hao Chi et al. Proposing new early detection indicators for pancreatic cancer: Combining machine learning and neural networks for serum mirna-based diagnostic model. *Frontiers in Oncology*, 13:1244578, 2023.

[119] Nikhil Cheerla and Olivier Gevaert. Microrna based pan-cancer diagnosis and treatment recommendation. *BMC bioinformatics*, 18(1):1–11, 2017.

[120] Yuanyuan Li et al. A comprehensive genomic pan-cancer classification using the cancer genome atlas gene expression data. *BMC genomics*, 18:1–13, 2017.

[121] Yuyan Xu, Wei Liao, Huanwei Chen, and Mingxin Pan. Constructing diagnostic signature of serum micrornas using machine learning for early pan-cancer detection. *Discover Oncology*, 15(1):263, 2024.

[122] Ruidong Li et al. Cancermirnome: an interactive analysis and visualization database for mirnome profiles of human cancer. *Nucleic acids research*, 50(D1):D1139–D1146, 2022.

[123] Baoxing Tian et al. A novel tcga-validated, mirna-based signature for prediction of breast cancer prognosis and survival. *Frontiers in Cell and Developmental Biology*, 9:717462, 2021.

[124] Shuting Lin et al. Integrative analysis of tcga data identifies mirnas as drug-specific survival biomarkers. *Scientific Reports*, 12(1):6785, 2022.

[125] Banabithi Bose, Matthew Moravec, and Serdar Bozdag. Computing microrna-gene interaction networks in pan-cancer using mirdriver. *Scientific Reports*, 12(1):3717, 2022.

[126] Raviprasad Kuthethur et al. An integrated analysis of micrornas regulating dna damage response in triple-negative breast cancer. *Breast Cancer*, 30(5):832–844, 2023.

[127] Sharif Moradi et al. Pan-cancer analysis of microrna expression profiles highlights micrornas enriched in normal body cells as effective suppressors of multiple tumor types: A study based on tcga database. *PloS one*, 17(4):e0267291, 2022.

[128] Alejandro Lopez-Rincon, Alberto Tonda, Mohamed Elati, Olivier Schwander, Benjamin Piwowarski, and Patrick Gallinari. Evolutionary optimization of convolutional neural networks for cancer mirna biomarkers classification. *Applied Soft Computing*, 65:91–100, 2018.

[129] Yifeng Li, Chih-Yu Chen, and Wyeth W Wasserman. Deep feature selection: theory and application to identify enhancers and promoters. *Journal of Computational Biology*, 23(5):322–336, 2016.

[130] Boyu Lyu and Anamul Haque. Deep learning based tumor type classification using gene expression data. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pages 89–96, 2018.

[131] Milad Mostavi, Yu-Chiao Chiu, Yufei Huang, and Yidong Chen. Convolutional neural network models for cancer type prediction based on gene expression. *BMC medical genomics*, 13(5):1–13, 2020.

[132] Shamveel Hussain Shah et al. Optimized gene selection and classification of cancer from microarray gene expression data using deep learning. *Neural Computing and Applications*, pages 1–12, 2020.

[133] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[134] Ananya Raghu, Anisha Raghu, and Jillian F Wise. Deep learning–based identification of tissue of origin for carcinomas of unknown primary using microrna expression: algorithm development and validation. *JMIR Bioinformatics and Biotechnology*, 5:e56538, 2024.

[135] Tanya Barrett et al. Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research*, 41(D1):D991–D995, 2012.

[136] John N Weinstein et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

[137] Amrita Kundu, Joginder Singh, Jayanta Kumar Pal, and Shubhra Sankar Ray. Predicting drug-resistant mirnas in cancer. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 12(1):6, 2022.

[138] Verónica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. Feature selection for high-dimensional data. *Progress in Artificial Intelligence*, 5(2):65–75, 2016.

[139] Gareth James. An introduction to statistical learning with applications in r, 2013.

[140] F. Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[141] Jiawei Han, Jian Pei, and Hanghang Tong. *Data mining: concepts and techniques*. Morgan kaufmann, 2022.

[142] Piyushkumar A Mundra and Jagath C Rajapakse. Svm-rfe with mrmr filter for gene selection. *IEEE transactions on nanobioscience*, 9(1):31–37, 2009.

[143] Manoranjan Dash and Huan Liu. Consistency-based search in feature selection. *Artificial intelligence*, 151(1-2):155–176, 2003.

[144] Alok Sharma, Seiya Imoto, Satoru Miyano, and Vandana Sharma. Null space based feature selection method for gene expression data. *International Journal of Machine Learning and Cybernetics*, 3(8):269–276, 2012.

[145] Lei Yu and Huan Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 856–863, 2003.

[146] Arunkumar Chinnaswamy and Ramakrishnan Srinivasan. Hybrid feature selection using correlation coefficient and particle swarm optimization on microarray gene expression data. In *IBICA*, pages 229–239. Springer, 2016.

[147] Chuan Luo, Sizhao Wang, Tianrui Li, Hongmei Chen, Jiancheng Lv, and Zhang Yi. Large-scale meta-heuristic feature selection based on bpso assisted rough hypercuboid approach. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[148] Liudmila Prokhorenkova et al. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.

[149] Xingyu Zhao, Wei Huang, Xiaowei Huang, Valentin Robu, and David Flynn. Baylime: Bayesian local interpretable model-agnostic explanations. In *Uncertainty in artificial intelligence*, pages 887–896. PMLR, 2021.

[150] Scott M Lundberg et al. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

[151] Joginder Singh, Shubhra Sankar Ray, and Sukriti Roy. Identifying pan-cancer and cancer subtype mirnas using interpretable convolutional neural network. *Journal of Computational Science*, 85:102510, 2025.

[152] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[153] Jun-Hao Li et al. starbase v2. 0: decoding mirna-cerna, mirna-ncrna and protein–rna interaction networks from large-scale clip-seq data. *Nucleic acids research*, 42(D1):D92–D97, 2014.

[154] Kuan-Li Wu, Ying-Ming Tsai, Chi-Tun Lien, Po-Lin Kuo, and Jen-Yu Hung. The roles of microrna in lung cancer. *International Journal of Molecular Sciences*, 20(7), 2019.

[155] Ramesh Singh and Yin-Yuan Mo. Role of microrna s in breast cancer. *Cancer biology & therapy*, 14(3):201–212, 2013.

[156] Katey SS Enfield, Greg L Stewart, Larissa A Pikor, Carlos E Alvarez, Stephen Lam, Wan L Lam, and Raj Chari. Microrna gene dosage alterations and drug response in lung cancer. *Journal of Biomedicine and Biotechnology*, 2011(1):1–15, 2011.

[157] Kazuhiro Kitamura, Masahiro Seike, Tetsuya Okano, Kuniko Matsuda, Akihiko Miyanaga, Hideaki Mizutani, Rintaro Noro, Yuji Minegishi, Kaoru Kubota, and Akihiko Gemma. Mir-134/487b/655 cluster regulates tgf-β–induced epithelial–mesenchymal transition and drug resistance to gefitinib by targeting magi2 in lung adenocarcinoma cellsemt and resistance to gefitinib by mir-134/487b/655 cluster. *Molecular cancer therapeutics*, 13(2):444–453, 2014.

[158] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, 24(7):881–892, 2002.

[159] Richard Jensen and Qiang Shen. Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches. *IEEE Transactions on knowledge and data engineering*, 16(12):1457–1471, 2004.

[160] George J Klir and Bo Yuan. Fuzzy sets and fuzzy logic: theory and applications. *Possibility Theory versus Probab. Theory*, 32(2):1–5, 1996.

[161] Debashis Sen and Sankar K Pal. Generalized rough sets, entropy, and image ambiguity measures. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):117–128, 2008.

[162] Jayanta Kumar Pal, Shubhra Sankar Ray, Sung-Bae Cho, and Sankar K. Pal. Fuzzy-rough entropy measure and histogram based patient selection for mirna ranking in cancer. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(2):659–672, 2018.

[163] Didier Dubois and Henri Prade. Rough fuzzy sets and fuzzy rough sets. *International Journal of General System*, 17(2-3):191–209, 1990.

[164] Sampa Misra and Shubhra Sankar Ray. Finding optimum width of discretization for gene expressions using functional annotations. *Computers in biology and medicine*, 90:59–67, 2017.

[165] Tian Yang, Yuan-Jiang Li, Yuhua Qian, and Fei-Yue Wang. Consistent matrix: A feature selection framework for large-scale data sets. *IEEE Transactions on Fuzzy Systems*, 2023.

[166] John C Platt. 12 fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods*, pages 185–208, 1999.

[167] Qiancheng Song et al. mir-483-5p promotes invasion and metastasis of lung adenocarcinoma by targeting rhogdi1 and alcam. *Cancer research*, 74(11):3031–3042, 2014.

[168] Y Chen et al. Mir-513a-3p inhibits emt mediated by hoxb7 and promotes sensitivity to cisplatin in ovarian cancer cells. *Eur Rev Med Pharmacol Sci*, 24(20):10391–10402, 2020.

[169] Hong Wu, Hong-Yan Wei, and Qian-Qian Chen. Long noncoding rna hottip promotes the metastatic potential of ovarian cancer through the regulation of the mir-615-3p/smarce1 pathway. *The Kaohsiung journal of medical sciences*, 36(12):973–982, 2020.

[170] Lei Zhang et al. mir-153 supports colorectal cancer progression via pleiotropic effects that enhance invasion and chemotherapeutic resistancemir-153 supports colorectal cancer progression. *Cancer research*, 73(21):6435–6447, 2013.

[171] Yajing Zhang et al. Pan-cancer analysis based on epor expression with potential value in prognosis and tumor immunity in 33 tumors. *Frontiers in Oncology*, 12:844794, 2022.

[172] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[173] Gabriel Erion et al. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature machine intelligence*, 3(7):620–631, 2021.

[174] Geoffrey E Hinton et al. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[175] Utomo Pujianto et al. K-nearest neighbor (k-nn) based missing data imputation. In *2019 5th International Conference on Science in Information Technology (ICSITech)*, pages 83–88. IEEE, 2019.

[176] Siyuan Zheng et al. Comprehensive pan-genomic characterization of adrenocortical carcinoma. *Cancer Cell*, 29(5):723–736, 2016.

[177] The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, 507(7492):315–322, 2014.

[178] A. Gordon Robertson et al. Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell*, 171(3):540–556.e25, 2017.

[179] The Cancer Genome Atlas Research Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[180] The Cancer Genome Atlas Research Network. Integrated genomic and molecular characterization of cervical cancer. *Nature*, 543(7645):378–384, 2017.

[181] Farshad Farshidfar et al. Integrative genomic analysis of cholangiocarcinoma identifies distinct idh-mutant molecular profiles. *Cell Reports*, 18(11):2780–2794, 2017.

[182] Caleb F Davis et al. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell*, 26(3):319–330, September 2014.

[183] The Cancer Genome Atlas Research Network. Integrated genomic characterization of oesophageal carcinoma. *Nature*, 541(7636):169–175, 2017.

[184] The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, 513(7517):202–209, 2014.

[185] The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008.

[186] Cameron W. Brennan. The somatic genomic landscape of glioblastoma. *Cell*, 155(2):462–477, 2013.

[187] The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, 517(7536):576–582, 2015.

[188] Adrian Ally et al. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell*, 169(7):1327–1341.e23, 2017.

[189] Caleb F. Davis et al. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell*, 26(3):319–330, 2014.

[190] The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of papillary renal-cell carcinoma. *N. Engl. J. Med.*, 374(2):135–145, January 2016.

[191] The Cancer Genome Atlas Research Network. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N. Engl. J. Med.*, 372(26):2481–2498, June 2015.

[192] The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543–550, July 2014.

[193] Joshua D Campbell et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat. Genet.*, 48(6):607–616, June 2016.

[194] Anthony Tubbs and André Nussenzweig. Endogenous dna damage as a source of genomic instability in cancer. *Cell*, 168(4):644–656, 2017.

[195] Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, June 2011.

[196] Cancer Genome Atlas Research Network. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell*, 32(2):185–203.e13, August 2017.

[197] Lauren Fishbein et al. Comprehensive molecular characterization of pheochromocytoma and paraganglioma. *Cancer Cell*, 31(2):181–193, February 2017.

[198] Cancer Genome Atlas Research Network. The molecular taxonomy of primary prostate cancer. *Cell*, 163(4):1011–1025, November 2015.

[199] Cancer Genome Atlas Research Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, 2012.

[200] Adam Abeshouse et al. Comprehensive and integrated genomic characterization of adult soft tissue sarcomas. *Cell*, 171(4):950–965.e28, November 2017.

[201] Cancer Genome Atlas Network. Genomic classification of cutaneous melanoma. *Cell*, 161(7):1681–1696, June 2015.

[202] Hui Shen et al. Integrated molecular characterization of testicular germ cell tumors. *Cell Rep.*, 23(11):3392–3406, June 2018.

[203] Milan Radovich et al. The integrated genomic landscape of thymic epithelial tumors. *Cancer Cell*, 33(2):244–258.e10, February 2018.

[204] Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. *Cell*, 159(3):676–690, October 2014.

[205] Andrew D Cherniack et al. Integrated molecular characterization of uterine carcinosarcoma. *Cancer Cell*, 31(3):411–423, March 2017.

[206] A Gordon Robertson et al. Integrative analysis identifies four molecular and clinical subsets in uveal melanoma. *Cancer Cell*, 32(2):204–220.e15, August 2017.

[207] Naofumi Asano, Juntaro Matsuzaki, Makiko Ichikawa, Junpei Kawauchi, Satoko Takizawa, Yoshiaki Aoki, Hiromi Sakamoto, Akihiko Yoshida, Eisuke Kobayashi, Yoshikazu Tanzawa, et al. A serum microrna classifier for the diagnosis of sarcomas of various histological subtypes. *Nature communications*, 10(1):1299, 2019.

[208] Shixiong Wu et al. A five-microrna signature predicts the prognosis in nasopharyngeal carcinoma. *Frontiers in Oncology*, 11:723362, 2021.

[209] Jin Xu et al. Reluplex made more practical: Leaky relu. In *2020 IEEE Symposium on Computers and communications (ISCC)*, pages 1–7. IEEE, 2020.

[210] Qiuyu Zhu et al. Improving classification performance of softmax loss function based on scalable batch-normalization. *Applied Sciences*, 10(8):2950, 2020.

[211] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[212] Nitish Srivastava et al. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[213] Jonghoon Jin, Aysegul Dundar, and Eugenio Culurciello. Flattened convolutional neural networks for feedforward acceleration. *arXiv preprint arXiv:1412.5474*, 2014.

[214] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2):1929–1958, 2012.

[215] Mohammed Gamal Ragab et al. An ensemble one dimensional convolutional neural network with bayesian optimization for environmental sound classification. *Applied Sciences*, 11(10):4660–4679, 2021.

[216] Louis Owen. *Hyperparameter Tuning with Python: Boost your machine learning model's performance via hyperparameter tuning.* Packt Publishing Ltd, 2022.

[217] Rahul Chauhan, Kamal Kumar Ghanshala, and RC Joshi. Convolutional neural network (cnn) for image detection and recognition. In *2018 first international conference on secure cyber computing and communication (ICSCCC)*, pages 278–282. IEEE, 2018.

[218] Tanoy Debnath et al. Four-layer convnet to facial emotion recognition with minimal epochs and the significance of data diversity. *Scientific Reports*, 12(1):6991, 2022.

[219] Ramaprasad Poojary and Akul Pai. Comparative study of model optimization techniques in fine-tuned cnn models. In *2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, pages 1–4. IEEE, 2019.

[220] Sajid Ali et al. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99:101805, 2023.

[221] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[222] Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9269–9278, Vienna, Austria, 13–18 Jul 2020. PMLR.

[223] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

[224] Yuhao Chen and Xiaowei Wang. mirdb: an online database for prediction of functional microrna targets. *Nucleic acids research*, 48(D1):D127–D131, 2020.

[225] National Human Genome Research Institute. Gene Ontology Analysis, 2023. `https://www.geneontology.org/`, Last accessed on May-2023.

[226] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[227] Wei Li et al. Circular rna circ-ccac1 facilitates adrenocortical carcinoma cell proliferation, migration, and invasion through regulating the mir-514a-5p/c22orf46 axis. *BioMed Research International*, 2020(1):3501451, 2020.

[228] Luca Falzone et al. Prognostic significance of deregulated microrna s in uveal melanomas. *Molecular medicine reports*, 19(4):2599–2610, 2019.

[229] Bao Zang, Jianqiang Zhao, and Chen Chen. Lncrna pcat-1 promoted escc progression via regulating anxa10 expression by sponging mir-508-3p. *Cancer management and research*, pages 10841–10849, 2019.

[230] Hua Tan et al. Pan-cancer analysis on microrna-associated gene activation. *EBioMedicine*, 43:82–97, 2019.

[231] Chenming Zhong et al. Mir4435-2hg is a potential pan-cancer biomarker for diagnosis and prognosis. *Frontiers in Immunology*, 13:855078, 2022.

[232] Ganglin Su et al. Ythdf2 is a potential biomarker and associated with immune infiltration in kidney renal clear cell carcinoma. *Frontiers in pharmacology*, 12:709548, 2021.

[233] Ping Li et al. Downregulation of mirna-141 in breast cancer cells is associated with cell migration and invasion: involvement of anp32e targeting. *Cancer medicine*, 6(3):662–672, 2017.

[234] Teresa Bellissimo et al. Thymic epithelial tumors phenotype relies on mir-145-5p epigenetic regulation. *Molecular cancer*, 16(1):1–15, 2017.

[235] Carmen ON Leung et al. mir-135a leads to cervical cancer cell transformation through regulation of β-catenin via a siah1-dependent ubiquitin proteosomal pathway. *Carcinogenesis*, 35(9):1931–1940, 2014.

[236] Jacopo Manso et al. Overexpression of mir-375 and l-type amino acid transporter 1 in pheochromocytoma and their molecular and functional implications. *International Journal of Molecular Sciences*, 23(5):2413, 2022.

[237] C-J Wang, H Zou, and G-F Feng. Mir-10b regulates the proliferation and apoptosis of pediatric acute myeloid leukemia through targeting hoxd10. *European Review for Medical & Pharmacological Sciences*, 22(21), 2018.

[238] Junchao Huang et al. microrna mir-10b inhibition reduces cell proliferation and promotes apoptosis in non-small cell lung cancer (nsclc) cells. *Molecular bioSystems*, 11(7):2051–2059, 2015.

[239] Ju Cheol Son et al. mir-10a and mir-204 as a potential prognostic indicator in low-grade gliomas. *Cancer informatics*, 16:1176935117702878, 2017.

[240] Kunhao Wang et al. Microrna and gene networks in human diffuse large b-cell lymphoma. *Oncology letters*, 8(5):2225–2232, 2014.

[241] Laura Gonzalez dos Anjos et al. Could mirna signatures be useful for predicting uterine sarcoma and carcinosarcoma prognosis and treatment? *Cancers*, 10(9):315, 2018.

[242] Beatriz Nunes Schiavon et al. mirnas 144-3p, 34a-5p, and 206 are a useful signature for distinguishing uterine leiomyosarcoma from other smooth muscle tumors. *Surgical and Experimental Pathology*, 2:1–8, 2019.

[243] Yebin Lu et al. Mir-135a-5p suppresses trophoblast proliferative, migratory, invasive, and angiogenic activity in the context of unexplained spontaneous abortion. *Reproductive Biology and Endocrinology*, 20(1):82, 2022.

[244] Arkadiusz others Gajek. Current implications of micrornas in genome stability and stress responses of ovarian cancer. *Cancers*, 13(11):2690, 2021.

[245] Mohammad Reza Golbakhsh et al. Down-regulation of microrna-182 and microrna-183 predicts progression of osteosarcoma. *Archives of Medical Science*, 13(6):1352–1356, 2017.

[246] Ting-Hua Yan et al. Prognostic significance of long non-coding rna pcat-1 expression in human hepatocellular carcinoma. *International journal of clinical and experimental pathology*, 8(4):4126, 2015.

[247] Faezeh Malakoti et al. Dna repair and damage pathways in mesothelioma development and therapy. *Cancer Cell International*, 22(1):176, 2022.

[248] Solomon L. Woldu, James F. Amatruda, and Aditya Bagrodia. Testicular germ cell tumor genomics. *Current Opinion in Urology*, 27(1):41–47, 2017.

[249] Leimarembi Devi Naorem, Mathavan Muthaiyan, and Amouda Venkatesan. Identification of dysregulated mirnas in triple negative breast cancer: a meta-analysis approach. *Journal of cellular physiology*, 234(7):11768–11779, 2019.

[250] Radoslaw Charkiewicz et al. mirna-seq tissue diagnostic signature: A novel model for nsclc subtyping. *International Journal of Molecular Sciences*, 24(17):13318, 2023.

[251] Guanghui Ying et al. Identification of eight key mirnas associated with renal cell carcinoma: A meta-analysis. *Oncology letters*, 16(5):5847–5855, 2018.

[252] Jianwen Yu et al. Intrarenal microrna signature related to the fibrosis process in chronic kidney disease: identification and functional validation of key mirnas. *BMC nephrology*, 20(1):1–13, 2019.

[253] Gan Yu et al. mirna-34a suppresses cell proliferation and metastasis by targeting cd44 in human renal carcinoma cells. *The Journal of urology*, 192(4):1229–1237, 2014.

[254] Feng Xiong et al. Mir-204 inhibits the proliferation and invasion of renal cell carcinoma by inhibiting rab22a expression. *Oncology Reports*, 35(5):3000–3008, 2016.

[255] Christoph Molnar. Interpretable machine learning, 2020. `https://christophm.github.io/interpretable-ml-book/`, Last accessed on 2024-08-25.

[256] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[257] Babak Alipanahi et al. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.

[258] Hanjun Dai et al. Sequence2vec: a novel embedding approach for modeling transcription factor binding affinity landscape. *Bioinformatics*, 33(22):3575–3583, 2017.

[259] Yu Li et al. Deepre: sequence-based enzyme ec number prediction by deep learning. *Bioinformatics*, 34(5):760–769, 2018.

[260] Emad Dabous, Adel Guirgis, and Hany Khalil. Targeting tumor suppressor genes by mir-141 family as a potential regulatory function in cervical cancer. *Journal of Internal Medicine: Science & Art*, 4:47–53, 2023.

[261] Xianxu Yang and Ping Wang. Mir-188-5p and mir-141-3p influence prognosis of bladder cancer and promote bladder cancer synergistically. *Pathology-Research and Practice*, 215(11):152598, 2019.

[262] Shunji Tamagawa et al. Role of mir-200c/mir-141 in the regulation of epithelial-mesenchymal transition and migration in head and neck squamous cell carcinoma. *International journal of molecular medicine*, 33(4):879–886, 2014.

[263] Nasrin Zare et al. The expression level of hsa-mir-146a-5p in plasma-derived exosomes of patients with diffuse large b-cell lymphoma. *Journal of Research in Medical Sciences: The Official Journal of Isfahan University of Medical Sciences*, 24:1–7, 2019.

[264] Mariya Aksenenko et al. Differences in microrna expression between melanoma and healthy adjacent skin. *BMC dermatology*, 19(1):1–9, 2019.

[265] Soudeh Ghafouri-Fard, Mahdi Gholipour, and Mohammad Taheri. Microrna signature in melanoma: biomarkers and therapeutic targets. *Frontiers in Oncology*, 11:608987, 2021.

[266] Subhrajit Mitra, Rajarshi Mukhopadhyay, and Paramita Chattopadhyay. Pso driven designing of robust and computation efficient 1d-cnn architecture for transmission line fault detection. *Expert Systems with Applications*, 210:118178, 2022.

[267] Dusmurod Kilichev and Wooseong Kim. Hyperparameter optimization for 1d-cnn-based network intrusion detection using ga and pso. *Mathematics*, 11(17):3724, 2023.

[268] Sheikh Shanawaz Mostafa et al. Multi-objective hyperparameter optimization of convolutional neural network for obstructive sleep apnea detection. *IEEE Access*, 8:129586–129599, 2020.

[269] Jinguo Lyu, Taihong Hu, Guangwei Liu, Bo Cao, Wenqi Wang, and Shixu Li. Stability evaluation of open-pit mine slope based on bayesian optimization 1d-cnn. *Scientific Reports*, 14(1):13995, 2024.

[270] Lotfi A Zadeh. A note on z-numbers. *Information sciences*, 181(14):2923–2932, 2011.

[271] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[272] Gary C McDonald. Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):93–100, 2009.

[273] Amartya Sen. *On economic inequality*. Oxford university press, 1997.

[274] Jun-Hao Li, Shun Liu, Hui Zhou, Liang-Hu Qu, and Jian-Hua Yang. starbase v2.0: decoding mirna-cerna, mirna-ncrna and protein–rna interaction networks from large-scale clip-seq data. *Nucleic Acids Research*, 42(D1):D92–D97, 11 2013.

[275] Anna Barbato, Fabiola Piscopo, Massimiliano Salati, Luca Reggiani-Bonetti, Brunella Franco, and Pietro Carotenuto. Micro-rna in cholangiocarcinoma: implications for diagnosis, prognosis, and therapy. *Journal of Molecular Pathology*, 3(2):88–103, 2022.

[276] Roghayeh Sheervalilou, Amir Mahdi Khamaneh, Akbar Sharifi, Masoud Nazemiyeh, Ali Taghizadieh, Khalil Ansarin, and Nosratollah Zarghami. Using mir-10b, mir-1 and mir-30a expression profiles of bronchoalveolar lavage and sputum for early detection of non-small cell lung cancer. *Biomedicine & Pharmacotherapy*, 88:1173–1182, 2017.

[277] HL Ma, XP Wen, XZ Zhang, XL Wang, DL Zhao, SM Che, and CX Dang. mir-106a* inhibits the proliferation of esophageal carcinoma cells by targeting cdk2-associated cullin 1 (cacul1). *Cellular and Molecular Biology*, 61(4):56–62, 2015.

[278] Lin-hong Jiang, He-da Zhang, and Jin-hai Tang. Mir-30a: A novel biomarker and potential therapeutic target for cancer. *Journal of Oncology*, 2018:1–9, August 2018.

[279] Juan Cui, Joanna B Eldredge, Ying Xu, and David Puett. Microrna expression and regulation in human ovarian carcinoma cells by luteinizing hormone. *PLoS One*, 6(7):e21730, 2011.

[280] Elrasheid AH Kheirelseid, Nicola Miller, Kah Hoong Chang, Catherine Curran, Emer Hennessey, Margaret Sheehan, John Newell, Christophe Lemetre, Graham Balls, and Michael J Kerin. mirna expressions in rectal cancer as predictors of response to neoadjuvant chemoradiation therapy. *International journal of colorectal disease*, 28:247–260, 2013.

[281] Aurora Esquela-Kerscher and Frank J Slack. Oncomirs—micrornas with a role in cancer. *Nature reviews cancer*, 6(4):259–269, 2006.

[282] Shang-Gin Wu, Tzu-Hua Chang, Yi-Nan Liu, and Jin-Yuan Shih. Microrna in lung cancer metastasis. *Cancers*, 11(2), 2019.

[283] Gang Li, Fang Wu, Han Yang, Xia Deng, and Yawei Yuan. Mir-9-5p promotes cell growth and metastasis in non-small cell lung cancer through the repression of tgfbr2. *Biomedicine & Pharmacotherapy*, 96:1170–1178, 2017.

[284] Hua Fang, Yutao Liu, Yaohong He, Yang Jiang, Yaping Wei, Han Liu, Yueqing Gong, and Guangyu An. Extracellular vesicle-delivered mir-505-5p, as a diagnostic biomarker of early lung adenocarcinoma, inhibits cell apoptosis by targeting tp53aip1. *International journal of oncology*, 54(5):1821–1832, 2019.

[285] Wenhui Yang, Chengcheng Zhou, Mei Luo, Xuejiao Shi, Yuan Li, Zengmiao Sun, Fang Zhou, Zhaoli Chen, and Jie He. Mir-652-3p is upregulated in non-small cell lung cancer and promotes proliferation and metastasis by directly targeting lgl1. *Oncotarget*, 7(13):16703, 2016.

[286] Divya Kesanakurti, Dilip Rajasekhar Maddirela, Subramanyam Chittivelu, Jasti S Rao, and Chandramu Chetty. Suppression of tumor cell invasiveness and in vivo tumor growth by microrna-874 in non-small cell lung cancer. *Biochemical and biophysical research communications*, 434(3):627–633, 2013.

[287] Mohd Saif Zaman, Sobha Thamminana, Varahram Shahryari, Takeshi Chiyomaru, Guoren Deng, Sharanjot Saini, Shahana Majid, Shinichiro Fukuhara, Inik Chang, Sumit Arora, et al. Inhibition of pten gene expression by oncogenic mir-23b-3p in renal cancer. *PloS one*, 7(11):e50203, 2012.

[288] Julia Liep, Ergin Kilic, Hellmuth A Meyer, Jonas Busch, Klaus Jung, and Anja Rabien. Cooperative effect of mir-141-3p and mir-145-5p in the regulation of targets in clear cell renal cell carcinoma. *PloS one*, 11(6):e0157801, 2016.

[289] Liwen Zhao, Kaihao Liu, Xiang Pan, Jing Quan, Liang Zhou, Zuwei Li, Canbin Lin, Jinling Xu, Weijie Xu, Xin Guan, et al. mir-625-3p promotes migration and invasion and reduces apoptosis of clear cell renal cell carcinoma. *American journal of translational research*, 11(10):6475, 2019.

[290] Sanghak Han, Hua Zou, Jin-Won Lee, Jeonghee Han, Heung Cheol Kim, Jeong Jin Cheol, Lee-Su Kim, and Haesung Kim. mir-1307-3p stimulates breast cancer development and progression by targeting smyd4. *Journal of Cancer*, 10(2):441, 2019.

[291] Isabel Stückrath, Brigitte Rack, Wolfgang Janni, Bernadette Jäger, Klaus Pantel, and Heidi Schwarzenbach. Aberrant plasma levels of circulating mir-16, mir-107,

mir-130a and mir-146a are associated with lymph node metastasis and receptor status of breast cancer patients. *Oncotarget*, 6(15):13387, 2015.

[292] Q Li, Y Yao, G Eades, Z Liu, Y Zhang, and Q Zhou. Downregulation of mir-140 promotes cancer stem cell formation in basal-like early stage breast cancer. *Oncogene*, 33(20):2589–2600, 2014.

[293] Yuliang Pan, Jun Zhang, Huiqun Fu, and Liangfang Shen. mir-144 functions as a tumor suppressor in breast cancer through inhibiting zeb1/2-mediated epithelial mesenchymal transition process. *OncoTargets and therapy*, pages 6247–6255, 2016.

[294] Gillian Browne et al. Microrna-378-mediated suppression of runx1 alleviates the aggressive phenotype of triple-negative mda-mb-231 human breast cancer cells. *Tumor Biology*, 37:8825–8839, 2016.

[295] Fermín MAR-AGUILAR, Claudia M LUNA-AGUIRRE, J Claudio MORENO-ROCHA, Juan ARAIZA-CHÁVEZ, Victor Trevino, Cristina RODRÍGUEZ-PADILLA, and Diana RESÉNDEZ-PÉREZ. Differential expression of mir-21, mir-125b and mir-191 in breast cancer tissue. *Asia-Pacific Journal of Clinical Oncology*, 9(1):53–59, 2013.

[296] Shuang Wang et al. Identification of three circulating micrornas in plasma as clinical biomarkers for breast cancer detection. *Journal of Clinical Medicine*, 12(1):322, 2022.

[297] Chong Chen et al. Microrna-3613-3p functions as a tumor suppressor and represents a novel therapeutic target in breast cancer. *Breast Cancer Research*, 23:1–13, 2021.

[298] Ilya A Balabin and Richard S Judson. Exploring non-linear distance metrics in the structure–activity space: Qsar models for human estrogen receptor. *Journal of Cheminformatics*, 10:1–11, 2018.

[299] Dor Kedem, Stephen Tyree, Fei Sha, Gert Lanckriet, and Kilian Q Weinberger. Non-linear metric learning. *Advances in neural information processing systems*, 25, 2012.

[300] Rajarajeswari Balasubramaniyan et al. Clustering of gene expression data using a local shape-based similarity measure. *Bioinformatics*, 21(7):1069–1077, 2005.

[301] Ahmad Taher Azar. Overview of type-2 fuzzy logic systems. *International Journal of Fuzzy System Applications (IJFSA)*, 2(4):1–28, 2012.

[302] Angelo Gaeta, Vincenzo Loia, and Francesco Orciuoli. An explainable prediction method based on fuzzy rough sets, topsis and hexagons of opposition: Applications to the analysis of information disorder. *Information Sciences*, 659:120050, 2024.

[303] Tsung-Han Chang. Fuzzy vikor method: A case study of the hospital service evaluation in taiwan. *Information Sciences*, 271:196–212, 2014.

[304] Petra Murinová, Karel Fiala, Stefania Boffa, and Vilém Novák. Graded hexagon of opposition in fuzzy natural logic with new intermediate quantifiers. *International Journal of Approximate Reasoning*, page 109465, 2025.

[305] David E Goldberg. Genetic algorithm in search, optimization and machine learning, addison. *W esley Publishing Company, R eading, MA*, 1(98):9, 1989.

[306] James Kennedy, Russell C Eberhart, and Yuhui Shi. The particle swarm. *Swarm intelligence*, pages 287–325, 2001.

[307] Seyedali Mirjalili and Andrew Lewis. The whale optimization algorithm. *Advances in engineering software*, 95:51–67, 2016.

[308] Zainab Ali Syeda et al. Regulatory mechanism of microrna expression in cancer. *International journal of molecular sciences*, 21(5):1–18, 2020.