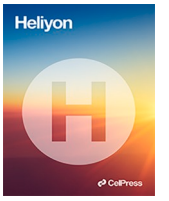
Heliyon 9 (2023) e17653

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



A deep learning-based framework for predicting survival-associated groups in colon cancer by integrating multi-omics and clinical data

Siamak Salimy [a], Hossein Lanjanian [b], Karim Abbasi [c], Mahdieh Salimi [d], Ali Najafi [e], Leili Tapak [f], Ali Masoudi-Nejad [a,*,1]

[a] Laboratory of System Biology and Bioinformatics (LBB), Department of Bioinformatics, University of Tehran, Kish International Campus, Kish, Iran
[b] Cellular and Molecular Endocrine Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[c] Laboratory of System Biology, Bioinformatics & Artificial Intelligent in Medicine (LBBai), Faculty of Mathematics and Computer Science, Kharazmi University, Tehran, Iran
[d] Department of Medical Genetics, Institute of Medical Biotechnology, National Institute of Genetic Engineering and Biotechnology (NIGEB), Tehran, Iran
[e] Molecular Biology Research Center, Systems Biology and Poisonings Institute, Tehran, Iran
[f] Department of Biostatistics, School of Public Health and Modeling of Noncommunicable Diseases Research Center, Hamadan University of Medical Sciences, Hamadan, Iran

ARTICLE INFO

Keywords:
Colorectal cancer
Data integration
Deep autoencoder
Gene expression
Multi-omics data
Precision medicine

ABSTRACT

Precise prognostic classification of patients and identifying survival subgroups and their associated genes can be important clinical references when designing treatment strategies for cancer patients. Multi-omics and data integration techniques are powerful tools to achieve this goal. This study aimed to introduce a machine learning method to integrate three types of biological data, and investigate the performance of two other methods, in identifying the survival dependency of patients. The data included TCGA RNA-seq gene expression, DNA methylation, and clinical data from 368 patients with colon cancer also we use an independent external validation data set, containing 232 samples. Three methods including, hyper-parameter optimized autoencoders (HPOAE), normal autoencoder, and penalized principal component analysis (PPCA) were used for simultaneous data integration and estimation under a COX hazards model. The HPOAE was thought to outperform other methods. The HPOAE had the Log Rank Mantel-Cox value of 14.27 $\pm$ 2, and a Breslow-Generalized Wilcoxon value of 13.13 $\pm$ 1. Ten miRNA, 11 methylated genes, and 28 mRNA all by (importance of marginal cutoff $>$ 0.95) were identified. The study demonstrated that hsa-miR-485-5p targets both ZMYM1 and tp53, the latter of which has been previously associated with cancer in numerous studies. Furthermore, compared to other methods, the HPOAE exhibited a greater capacity for identifying survival subgroups and the genes associated with them in patients with colon cancer. However, all of the results were obtained by computational methods, and clinical and experimental studies are needed to validate these results.

* Corresponding author. Laboratory of Systems Biology and Bioinformatics (LBB) Institute of Biochemistry and Biophysics University of Tehran, Tehran, Iran.
E-mail address: amasoudin@ut.ac.ir (A. Masoudi-Nejad).
[1] WWW:http://LBB.ut.ac.ir.

https://doi.org/10.1016/j.heliyon.2023.e17653
Received 30 April 2023; Received in revised form 30 May 2023; Accepted 25 June 2023
Available online 5 July 2023

1. Introduction

Colorectal cancer ranks third in terms of its incidence, following lung and breast cancer, and is the second most common cause of cancer-related deaths worldwide, as reported by the International Agency for Research on Cancer (IARC), the Global Cancer Observatory (GCO), and recent research [1]. This number is projected to reach approximately 1,919,534 by 2040 [2]. Recent reports indicate that the incidence rate of new cases of colorectal cancer is 38.2 per 100,000 men and women annually [3]. The death rate due to colorectal cancer is also high, at 13.7 per 100,000 men and women annually worldwide. These rates are based on 2013-2017 cases, adjusted for age, and the number of deaths that occurred in 2014–2018. In 2020, an estimated 104,610 new cases of colorectal cancer were reported, with 53,200 deaths recorded [4]. Based on 2015-2017 data, approximately 4.2% of people will be diagnosed with colorectal cancer at some point in their lifetime. In 2017, it was estimated that 1,348,087 individuals were living with this type of cancer in the United States [5]. Moreover, five-year survival rates of cancer patients are correlated with the stage of cancer and vary from 95% for stage I to 11% for stage IV. Early diagnosis of colon cancer could improve the chance of successful treatment for patients [6].

Treatment strategies for colon cancer are limited and unclear; thus, there is an urgent need to expand the number of tools to predict patient survival [7]. Numerous studies have been performed to identify the causative agents of cancer [8–10], the genomic and transcriptomic factors affecting it [11,12], and the diagnosis or treatment strategies [13–15]. To date, many types of data (genomics, transcriptomics, DNA methylation, etc.) have been examined in various studies to determine how they affect the survival of the patient for various cancer types. However, fusing these factors is still a challenge [16]. There are many methods available to integrate multi-omics data [17–19], including recently developed machine learning-based [20,21]. The introduction of new solutions, as well as new methods and roadmaps, indicates that the issue of data integration can still be considered important [22,23]. One of the concepts related to multi-omics data integration is survival analysis [24–26]. In survival analysis, the classification of patients is studied based on the level of risk for each patient, which is usually dependent on the stage of the disease [27–30], but there is a lack of stage-independent classification in colon cancer [31–33]. In our approach, the challenge of stage-independent classification is investigated. To this end, multi-omics data, which has been aggregated, and the dimension of data decremented by using a deep autoencoder, has been used [34]. In a deep autoencoder, an input is fed into the encoder, which embeds the input into a hidden representation, and then this representation is fed into a decoder which reconstructs the original input space [35]. During the training, the encoder learns a mapping of x to y encoding and the decoder then learns a nonlinear mapping from the encoded x into the origin space. The goal of training is to minimize reconstruction loss [36,37]. One of the major factors in machine learning-based methods is hyperparameter optimization, which reduces the human effort required to apply machine learning and improves the performance of the machine learning algorithms. This has resulted in new peak performance for key machine-learning benchmarks in several studies [38,39].

However, many studies ignored the optimization step in the pipeline, in similar work, Chaudhary studied survival in patients with liver cancer. In another study on this subject, survival in patients with colon cancer was examined [40], but the parameters of the deep learning [41] model are the same as in the Chaudhary model, and the optimizing of the parameters of the deep learning model is not presented. In all studies, principal component analysis (PCA) has been used as the first approach, but in our study, this model has been optimized with the penalized principal component analysis (PPCA) model. The presentation of the parameter optimization phase in the general pipeline, the introduction of the PPCA model for the first time in these studies, the use of a binary decision tree to create a predictive model, and the use of online web tools can all be considered as innovations of this study. It seems that the introduction of genes and groups related to survival, as well as the predictive platform introduced in this study, can have a good effect on the choice of an appropriate treatment method in patients with colon cancer.

One of the most common methods of feature selection is using PCA but studies have shown that this technique works by removing non-original components [42,43]. To cover the weak points of this method, the method of feature reduction using autoencoders is used. In addition to all the advantages of this method, one of its disadvantages is the presence of various hyperparameters, which are usually not considered in studies [44].

In classical feature reduction methods, some valuable biological data are ignored or truncated however this part of the data also could help to improve the predictive models. He goal of this study is to enhance the precision of the model and decrease the mean squared error (MSE), in order to improve the identification of survival-associated subgroups in individuals with colorectal cancer (CRC).To this aim in addition to using PCA and autoencoder, we optimized hyperparameters and produced a hyperparameter-optimized model of the autoencoder [45]. The study was conducted in two phases. Firstly, an unsupervised approach was used to create a deep-learning-based model that was optimized with hyperparameters. This model was trained using clinical data and multi-omics data to identify associations between omics features and patient survival. Secondly, a supervised approach was employed, which utilized the labels predicted during the unsupervised phase to create a rule-based model capable of extracting rules from multi-omics data. The results indicate that the model was effective in reducing the dimension of the input data. An online service was also developed to classify patients into high-risk and low-risk categories. Finally, certain miRNAs, methylated genes, and mRNAs were identified as being more closely linked to colon cancer.

2. Materials and methods

2.1. Datasets and study design

To conduct this study, we utilized two separate cohorts. First, we employed the multi-omics TCGA COAD dataset (n = 368) to obtain survival-risk class labels and train a rule-based, C5 pruning algorithm. Second, we utilized a confirmation dataset (n = 232) to evaluate the prediction accuracy of the deep learning-based model. Fig. 1 depicts the workflow of the study, which Fig. 1A, consists of an unsupervised section for identifying survival-associated subgroups and Fig. 1B a supervised section for predicting a model to identify survival-associated genes and discover rules between survival-associated subgroups.

2.2. Deep learning framework to transform the features

An autoencoder is a type of neural network that is unsupervised, feed-forward, and non-recurrent. Its purpose is to learn how to compress and encode data efficiently, then reconstruct the data from the encoded representation to a representation that is as close to the original input as possible [46].

Given an input layer taking the input ($x = x_1, \ldots, x_n$) of dimension n (here including miRNA, RNA and methylation data): In the autoencoder, the input, x, is transformed through the successive hidden layers to produce the output x' such that x' will be similar to x (note that x and x' have the same dimensions) [40]. For layer i, the rectified linear unit (Relu) as an activation function is as follows:

$$Y = f_i(x) = Relu(W_i x + b_i) \tag{1}$$

where x and y are two vectors of size d and p, respectively, W_i is the weight matrix of size (p), and b_i is an intercept vector of size p.

For an autoencoder with k layers, x' is then given by:

$$x' = F_{1\to k}(x) = f_1^{\circ} \ldots f_{k-1}^{\circ} f_k(x) \tag{2}$$

where $f_{k-1}^{\circ} f_k(x) = f_{k-1}(f_k(x))$ is the composing function of f_{k-1} with f_k. The goal of training an autoencoder is to find the weight vectors W_i of each layer by minimizing the reconstruction loss.

2.3. Hyper-parameter optimization

In machine learning, hyperparameter optimization, also known as tuning, refers to the process of selecting the most suitable set of hyperparameters for a learning algorithm. Hyperparameters are parameters used to regulate the learning process, while the values of other parameters, such as node weights, are learned through the algorithm. To implement an autoencoder, hyperparameters should be set. In this work, the hyper-parameter optimization step was set to find the best hyperparameters for a DL-base framework implementation. To this end, the Scikit-Optimize package was used [47]. These hyperparameters were the activation functions ReLu, sigmoid, linear, and tanh, with the optimizer functions Adam, SGD, Adagrad, Adadelta, RMSprop, Adama, and Nadam, and batch sizes of 32, 64,128,170, and 210.

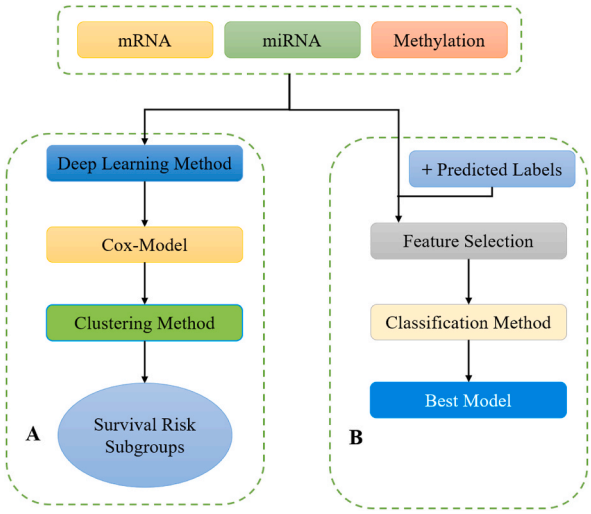


Fig. 1. The overall workflow of the study. A: The unsupervised section of the study to identify survival-associated subgroups. B: The supervised section to predict a model for identifying survival-associated Genes and discover rules between survival-associated subgroups.

2.4. Survival-associated features and segmentation

We reduced feature dimensionality using a DL-framework and constructed a multivariate Cox proportional hazards (Cox-PH) model for each transformed feature. Then, we employed these survival-associated features to cluster samples using the k-means clustering method. We determined the optimal cluster number using the Silhouette index [48], as implemented in the R survcomp package for k-means [49].

2.5. Data partitioning and robustness evaluation

The TCGA dataset samples were randomly divided into five folds using the R caret package. Three folds were designated as the training set, while the other two were set aside as the test set. This process was similar to cross-validation (CV), which was discussed in a previous study [38]. Ten new combinations (folds) were generated as a result. Each new combination's training set (70% of samples) was utilized to develop a model, which was then assessed on the corresponding test set (30% of samples). The model's robustness was assessed by calculating Precision, Recall, F1-score, and Accuracy.

3. Alternative approaches to the DL framework

The performance of our DL framework was compared with some base approaches. Recent studies have shown that comparable approaches perform better than principal component analysis [40]. In the first base approach, the penalized principal component analysis (PPCA) was used, and the second base approach was an unoptimized autoencoder in which the hyperparameters were set randomly. These two base approaches were compared to a hyperparameter-optimized autoencoder (an autoencoder in which the hyperparameters are optimized). For all approaches, the number of features is equal and Cox-PH analysis, K-means clustering process, and survival analysis are performed.

3.1. Confirmation cohort

We used the confirmation cohort (E-GEOD-17538) as an input to feed the deep autoencoder with the same parameters and Cox-PH model created and the same clustering method used to cluster samples.

3.2. Supervised classification

Based on survival analysis and group separation, we have performed a supervised classification. To this end, the machine learning-based feature selection [50] is done initially to select the top N label-related features and is then used to construct a C5 classification method. The workflow of this phase is shown in Fig. 2.

3.3. C5 conditional rule set

Based on the C5 pruning algorithm [51], we have built a model to infer a rule set for conditional statistics. It should be noted that 0 and 1 are assigned to high-risk and low-risk cases, respectively. The model was built based on three criteria that are represented by I

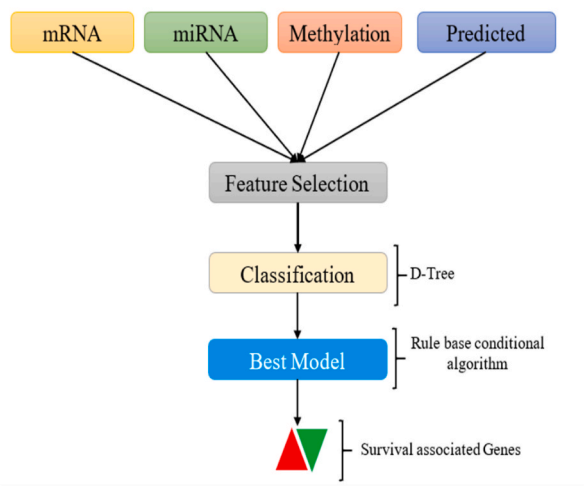


Fig. 2. The flowchart of the supervised phase of the study using identified labels from the previous section. In this phase, a model is predicted to discover genes that have a significant role and construct a rule-based model to select independent rules to classify patients.

to III as follows:

I: mRNA –base ruleset.

II: Methylate base ruleset.

III: miRNA –base ruleset.

4. Result

The description of Data is represented in (Table 1), and supplementary file 1.

The original data included samples (mRNA = 366, miRNA = 262, DNA methylation = 336) because missing some clinical data, 99 and 125 samples with mRNA and methylation are removed respectively (Table 1). shows the frequency distribution of the final samples selected for further analysis.

4.1. Transformed features using a DL framework

We utilized three preprocessed TCGA COAD omics datasets, comprising a total of 210 samples, as input for the autoencoder framework. To create a unique matrix, we stacked the three unit-norm scaled matrices by sample. The encoder structure consisted of two hidden layers, with 1000 and 100 nodes respectively, while the decoder structure had two hidden layers, one with 100 and another with 1000 nodes. We implemented the model using the Python Keras library available at https://github.com/fchollet/keras.The architecture of the proposed autoencoder is depicted in Fig. 3.

4.2. Hyper-parameter optimization

To calculate the optimal parameters, a permutation of all the elements (activation function, optimization function, and batch size) needs to be considered to calculate the optimal error rate such that the value of the mean square error (MSE) is minimized. For this, all the states for which these parameters could have been included in the model, and the value of the MSE, were calculated. It should be noted that the previously designed models did not select the optimal parameters. Therefore, the model is taught, based on three types of input omics data, two hidden and one dense layer, as well as a combination of said parameters for each model. For all models, the amount of MSE was calculated, and finally, the parameters providing the lower MSE rate were set as hyperparameters in the main model.

We repeated the whole path of implementation of the auto-encoder DL to see if the extracted features will result in the same labels (high/low risk survival groups). Kuder–Richardson criterion (takes values between 0 and 1 and greater values mean a better consistency) was used to evaluate consistency of the resulted clusters (0/1 for high/low risks) over 20 repetitions. We obtained a value of 0.912 indicating a strong consistency between clusters across repetitions.

In this study, activation functions, optimizer functions, and batch size were the hyper-parameters that were optimized to reduce the MSE between the input layer and output layer. It was shown that, after 1050 iterations, the best combination was Relu, RMSprop, and 32, with MSE = 0.015. The convergence plots are shown in Fig. 4.

4.3. Survival-associated features and segmentation

By DL-framework, the dimensions of features were reduced to 100 (the bottleneck layer), and features with a significant Cox-PH model were selected (log-rank $p < 0.05$) by the univariate Cox proportional hazards (Cox-PH) model. Based on the Silhouette index of the k-means clustering method, the optimal cluster number ($k = 2$) was selected, and samples were segmented into two groups.

4.4. Survival subtypes identified and clustered in TCGA multi-omics COAD data

From the TCGA COAD project, 211 samples that had RNA-Seq, miRNA-Seq, and DNA methylation data were obtained.

Samples were preprocessed and 16,596, 459, and 20,150 features were obtained for mRNA, miRNAs, and DNA methylation genes respectively. The features were projected into a low-dimensional hidden representation using an autoencoder DL framework. The dimension of the features was reduced to 100 and the bottleneck layer features were used as inputs to feed the univariate Cox-PH model. This model identified 80 survival-associated features with significant log-rank (log-rank $p < 0.05$). Samples with survival-associated features were clustered into two groups using the k-means algorithm with the C-index criterion. In Fig. 5, the Kaplan–Meier survival curves show the good separation of these groups. Then, these two groups were then labeled as low-risk and high-risk.

4.5. Confirmation cohort

The data set, E-GEOD-17538, contains 232 samples (A-AFFY-44, RNA-seq) where samples with RNA-Seq data were used as input data to the Dl-framework and a Cox-PH model was created using the same clustering method, and samples were assigned to low-risk and, high-risk groups that are clearly separated. The results are shown in Fig. 6.

Table 1
The descriptive table of the data used in this study. TCGA set-multi-omics colon adenocarcinoma data were obtained from the TCGA portal.

Data set	Datatype	Number of Samples	Age (Average ± sd	Platform
TCGA	mRNA-Seq	267 (F = 46.8%)	66.4 ± 13.1	UNC IlluminaHiSeq_RNASeqV2
TCGA	miRNA-Seq	262 (F = 37.8%)	65.8 ± 13.3	BCGSC IlluminaHiSeq_miRNASeq
TCGA	DNA methylation	211 (F = 46.4%)	66.7 ± 12.8	JHU-USC HumanMethylation27
TCGA	Clinical data	388 (F = 46.3%)	67 ± 12.7	TCGA Colon Cancer Clinical metadata
E-GEOD-17538[a]	RNA-Seq	267 (F = 41.2%)	66.4 ± 13.1	transcription profiling by array

[a] Independent External validation dataset.

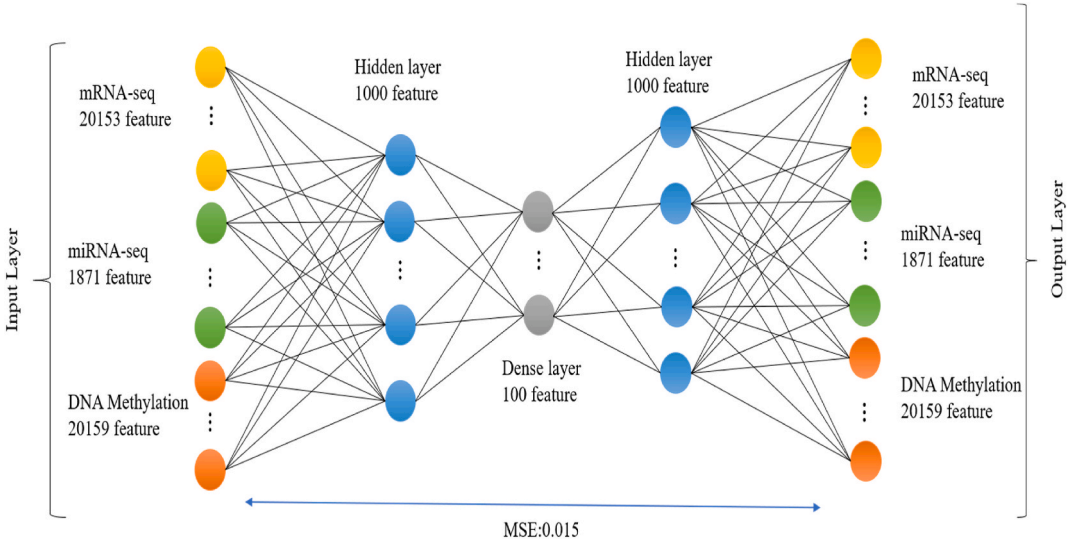


Fig. 3. The architecture of Autoencoder Deep Learning Framework. Three omics are imported as the input layer, and the projected features in the dense layer are used as new reduced features to create a CoxPH model to find survival-associated subgroups.

4.6. Alternative approaches to the DL framework

The penalized principal component analysis (PPCA) and an unoptimized autoencoder are used as alternative approaches to the DL framework. The comparative results are shown in (Table 2).

Comparing the survival function (Fig. 7A) and hazard function (Fig. 7B) across Normal Autoencoder, survival function (Fig. 7C) and hazard function (Fig. 7D) across Optimized Autoencoder, and survival function (Fig. 7E) and hazard function (Fig. 7F) across Penalized Principal Component, it is evident that there is a clear distinction between the two groups. Additionally, when examining the results of Cox analysis, it becomes possible to identify the groups that are associated with survival outcomes.

4.7. Supervised classification

Through the application of survival analysis and group separation, two distinct groups were identified and assigned labels denoting low and high risk. Using a machine learning-based approach to feature selection, the most important features were identified based on a likelihood ratio measure, and only those with a marginal cutoff ≥ 0.95 and an importance marginal cutoff exceeding 0.95 were selected. Specifically, this process resulted in the selection of 8778 mRNAs, 455 miRNAs, and 109 methylated genes as top features with significant associations.

We applied the C5 pruning algorithm to discover association rules between selected genes. The C5 machine learning, rule-based feature selection software is used to accurately select features which include 10 miRNA, 11 methylated genes, and 28 mRNA, (importance of marginal cutoff ≥ 0.95) for all, as shown in (Table 3).

Among 28 mRNA, 6 mRNA as miRNA targets were predicted by MIRWALK[2], including ACTL10, CAMK2B, CCL24, FAM122C, LBP, ZMYM1, and ZMYM1 reported as a mirTarBase[3] validated gene.

(Table 4) presents the target prediction scores obtained from two distinct web-based prediction tools.

[2] http://mirwalk.umm.uni-heidelberg.de.
[3] https://mirtarbase.cuhk.edu.cn.

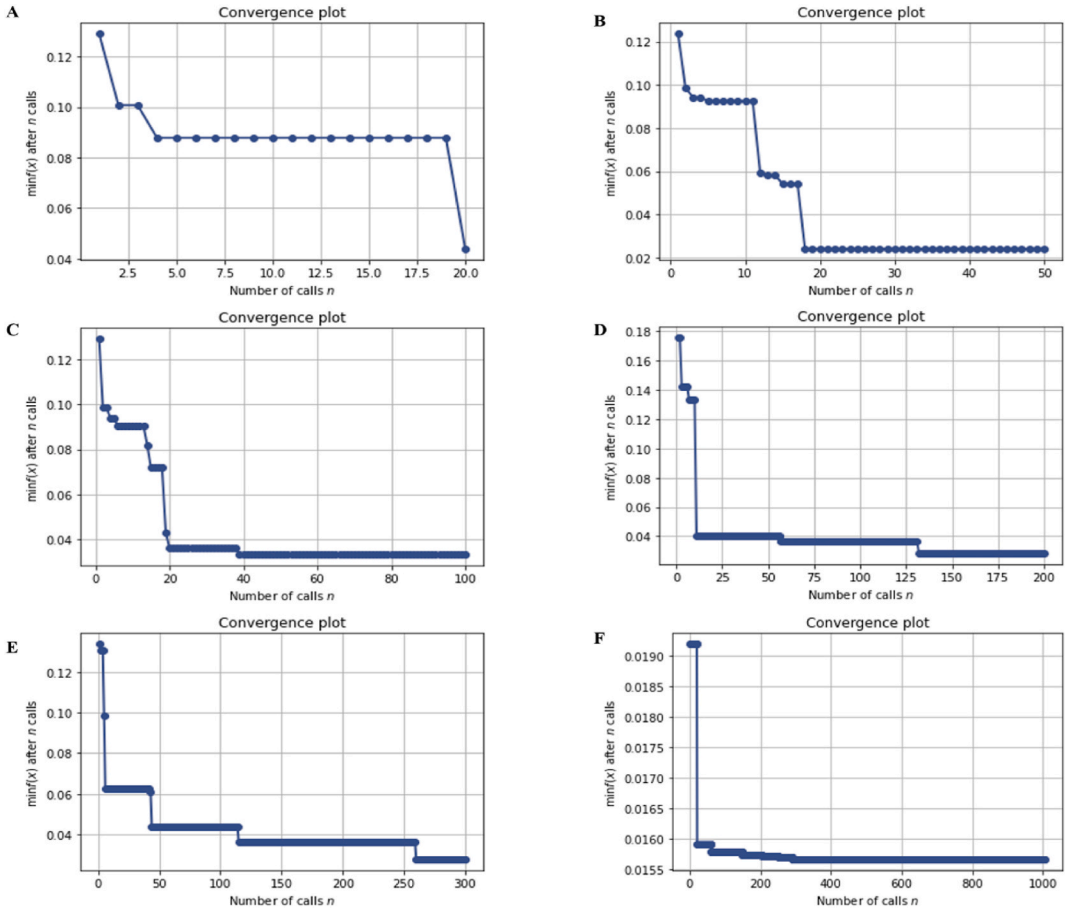


Fig. 4. Convergence plot after 20, 50, 100, 200, 300, and 1000 iterations from (A) to (F) respectively. As shown, after 20 iterations, MSE reached the minimum amount, and increasing the number of repetitions is ineffective.

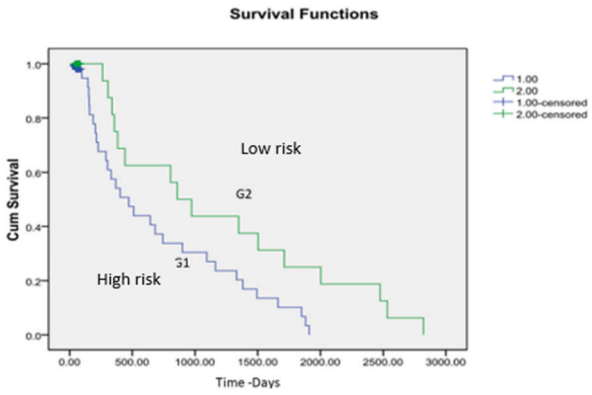


Fig. 5. Survival Function, the Kaplan – Meier survival curve. The Blue line indicates the high-risk group, and the green line indicates the low-risk group. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Two prediction tools, TargetScan[4] and miR-MicroT, were employed to identify target genes for miR-6842 and miR-33b-3p. The TargetScan Context++ score, which provides real-time PCR predictions, was used to indicate a higher probability of targeting, while a more negative MR-microT and miTG score was used to indicate greater suppression effects. Predicted miRNA and common targets for

[4] https://www.targetscan.org.

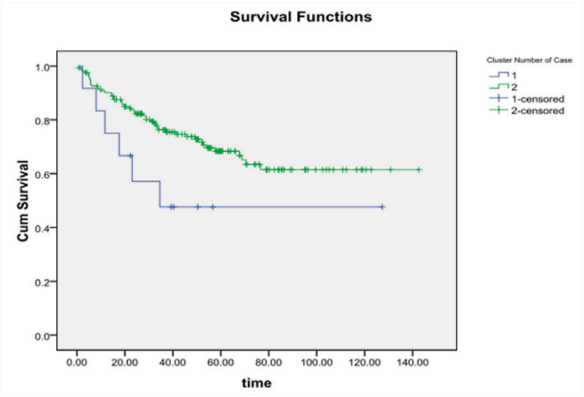


Fig. 6. The survival function of the confirmation cohort. As shown, two groups are separated based on the Dl-frame work and Cox-PH model that trained with the TCGA cohort.

Table 2
The comparison of the two base approaches includes penalized multivariate analysis and normal autoencoder with an optimized autoencoder. Optimized Autoencoder outperforms the other approaches.

Model	LogRank (Mantel-Cox) Chi-Square	Breslow (Generalized Wilcoxon) Chi-Square	Tarone-are Chi-Square	Sig
Optimized Auto encoder	14.272	12.012	13.131	0.001
Penalized Multivariate Analysis	9.914	2.105	5.761	0.002
Normal Autoencoder	8.375	2.743	4.49	0.004

every two miRNAs are shown in (Table 5). It may be concluded that both miRNAs may be indirectly related to each other because they both target the same gene. Of course, this issue should be investigated by other studies or researched experimentally.

4.8. Methylated genes

The research conducted in this study involved an examination and analysis of the field of epigenetics. Based on this analysis, it was concluded that methylation in certain genes, including ALG13, ARHGAP23, CAPS, DCT, DHDDS, E2F2, EPB41L1, GPR182, HEATR9, IGSF6, and ARHGEF37, may have a significant impact on patient survival. These genes were identified through predictions made using the targetscan and mirDB databases. Additionally, the methylation in the ANGPT4 gene was validated using the mirTarbase database.

The results of the study show the correlation between colon cancer and gene methylation for certain methylated genes, which have been identified as predicted targets of ten specific miRNAs. The significance of this correlation is expressed through the *P*-values of the individual genes, as follows: ALG13 with a *P*-value of 1.856e-10, ANGPT4 with a *P*-value of 8.364e-01, ARHGAP23 with a *P*-value of 5.557e-09, ARHGEF37 with a *P*-value of 1.017e-11, CAPS with a *P*-value of 1.582e-03, DCT with a *P*-value of 5.698e-03, DHDDS with a *P*-value of 1.536e-05, E2F2 with a *P*-value of 2.082e-11, EPB41L1 with a *P*-value of 1.624e-12, GPR182 with a *P*-value of 1.938e-03, and HEATR9 with a *P*-value of 2.024e-12.

4.9. Robustness and reproducibility evaluation

To validate the robustness of the two inferred survival risk groups obtained by the autoencoder, a classification model was built using the C5 classifier with CV (Fig. 2). TCGA samples were randomly separated into training (70%) and test (30%) sets (Table 6). shows a high Precision (0.978 ± 0.02), Recall (0.9375 ± 0.07), F1-score (0.9375 ± 0.03) and Accuracy (0.969 ± 0.01) On average, the training set consisting of three-omics produced comparable outcomes. The test data also demonstrated similar findings, with Precision of 0.56 ± 0.32, Recall of 0.56 ± 0.32, F1-score of 0.55 ± 0.30, and Accuracy of 0.71 ± 0.1. Regarding the testing of each omics dataset, this model produced notable but slightly less impressive outcomes, as shown in (Table 6). These results confirmed that the two inferential survival risk groups were strong enough to withstand the stochastic processes inherent in automatic encoder construction and training sample selection. It was found that using multiomics data was more effective in constructing the model compared to single-omics data.

As mentioned in this study, we were able to introduce various biomarkers based on genomics and epigenomics Data. Shown in Fig. 8A are the miRNAs that were discovered through our study and those that have been experimentally validated by the Mirwalk database. Fig. 8B introduces the mRNA obtained by our study with the experimentally validated mRNA based on the Mirwalk database in order to better understand the relationship between the two. Because of this, we can highlight the novel connection between the elements that this study found to be shared and that was experimentally confirmed by other studies, the Mirwalk, and the mirTarbase

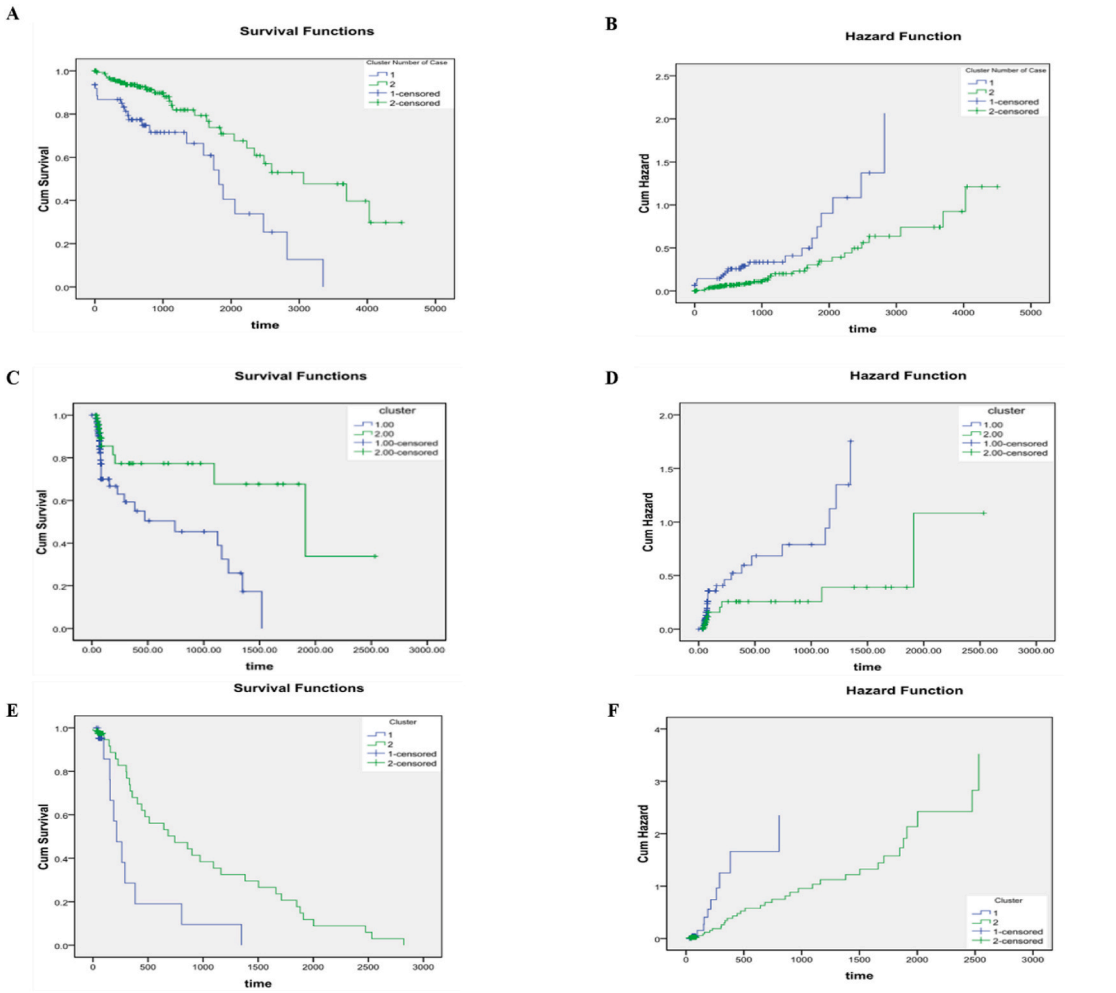


Fig. 7. Comparison of Survival Function, and Hazard Function on Normal Autoencoder (A, B), optimized autoencoder (C, D), and Penalized principal component (E, F). In all cases, two groups are separated clearly, and by considering the results of Cox analysis, survival-related groups could be identified.

Table 3
The ML-Base features selected from multi-omics (mRNA, miRNA, Methylation) data.

mRNA	mRNA	miRNA	Methylated gene
ACTL10	KCNC1	hsa-miR-3130-5p	ALG13
ARSH	KCNV2	hsa-miR-3189-5p	ANGPT4
BORCS7	KNG1	hsa-miR-33b-3p	CAPS
CAMK2B	LBP	hsa-miR-34a-5p	DHDDS
CCL24	LRRC69	hsa-miR-365b-5p	E2F2
CFAP52	NTN4	hsa-miR-3940-5p	GPR182
CYP11A1	PPM1D	hsa-miR-424-5p	HEATR9
CYP46A1	RASSF4	hsa-miR-485-5p	ARHGAP23
FAM122C	RSPH10B2	hsa-miR-6842-5p	ARHGEF37
FMO3	SLC2A14	hsa-miR-7704	DCT
FRA10AC1	ZG16		EPB41L1
HDGFL3	ZMYM1		IGSF6
HLA-DOB			

databases. Fig. 8C. The intricate relationship between miRNAs and the mRNA network discovered through the study is depicted in Fig. 9, Fig. 10, and vice versa. It should be noted that the mirwalk database with a cutoff of 0.95 and mirTarbase validation served as the basis for all of these items.

As we can see in the networks shown in Figs. 9 and 10, the relationship between miRNA and mRNA obtained by our study and those

Table 4
The target prediction scores of two popular web application prediction tools, Targetscan and MR-microT, were compared in this study for the prediction of target genes of miR-6842 and miR-33b-3p. The Targetscan Context++ score, which provides almost real-time PCR predictions, showed a higher value, indicating a greater likelihood of targeting. On the other hand, the MR-microT miTG score had a more negative value, suggesting a greater suppression effect.

		targetscan Context++ score	MR-microT miTG score
hsa-miR-6842-5p	ACTL10	−0.54	0.76055952
hsa-miR-33b-3p	LBP	−0.24	0.883622137

Table 5
Predicted miRNA and common target for every two miRNAs, it may be interpreted that both miRNAs may be indirectly related to each other because they both target the same gene. Of course, this issue should be investigated by other studies or researched experimentally.

miRNA ID	Target gene	miRNA ID	Target gene	miRNA ID	Target gene
hsa-miR-33b-5p	AGO2	hsa-miR-6842-5p	CDKN1A	hsa-miR-485-5p	QSOX1
hsa-miR-3940-5p		hsa-miR-7704		hsa-miR-6842-5p	
hsa-miR-485-5p	APOL6	hsa-miR-34a-5p	DFFA	hsa-miR-3130-5p	RAB1B
hsa-miR-7704		hsa-miR-6842-5p		hsa-miR-6842-5p	
hsa-miR-424-5p	ARHGDIA	hsa-miR-34a-5p	EFNB1	hsa-miR-3130-5p	SF3B3
hsa-miR-485-5p		hsa-miR-6842-5p		hsa-miR-34a-5p	
hsa-miR-34a-5p	CCL22	hsa-miR-34a-5p	GDE1	hsa-miR-424-5p	SLC39A9
hsa-miR-485-5p		hsa-miR-7704		hsa-miR-485-5p	
hsa-miR-3940-5p	TXNIP	hsa-miR-365b-5p	KLHL21	hsa-miR-3130-5p	STAT2
hsa-miR-424-5p		hsa-miR-485-5p		hsa-miR-485-5p	

Table 6
The 10-fold cross -validation is used to evaluate the C5 classifier's performance.

Dataset	10 - fold CV	Precision	Recall	F1-score	Accuracy
Training	3 - Omic	0.978 ± 0.02	0.9375 ± 0.07	0.9375 ± 0.03	0.969 ± 0.01
	mRNA Only	0.98 ± 0.02	0.94 ± 0.08	0.95 ± 0.04	0.97 ± 0.02
	miRNA Only	0.96 ± 0.03	0.93 ± 0.09	0.94 ± 0.05	0.96 ± 0.02
	Methylation Only	0.95 ± 0.04	0.88 ± 0.13	0.91 ± 0.07	0.94 ± 0.02
Test	3 - Omic	0.56 ± 0.32	0.56 ± 0.32	0.55 ± 0.30	0.71 ± 0.1
	mRNA Only	0.49 ± 0.31	0.50 ± .033	0.49 ± 0.32	0.65 ± .10
	miRNA Only	0.51 ± 0.32	0.53 ± 0.35	0.51 ± 0.32	0.66 ± .011
	Methylation Only	0.84 ± 0.12	0.85 ± 0.05	0.85 ± 0.03	0.82 ± 0.07

validated by databases and experimentally indicate that these biomarkers can be considered in the survival of patients with Colon cancer.

Overall, the studies suggest that the ZMYM1 gene may serve as a potential biomarker for predicting the progression and prognosis of colon cancer. Further research is needed to better understand the role of ZMYM1 in colon cancer and to validate its potential as a biomarker for clinical use. If validated, ZMYM1 could be used in combination with other biomarkers to improve the accuracy of colon cancer diagnosis and prediction of patient outcomes, ultimately leading to better patient care and outcomes.

By using the C5 D-tree, we discovered conditional rules and designed a conditional calculator that is available at the following address: (https://github.com/LBBSoft/ConditionalCalculator and 78.39.204.163/ConditionalCalculator). Instead of checking a large number of genes to determine the patient's risk (High or Low), the ConditionalCalculator platform can assist the doctor, researcher, or any other interested party to check only the biomarkers obtained from this research in the patient.

5. Discussion

CRC is a highly prevalent type of cancer worldwide, and the identification of biomarkers for the early detection and prediction of its progression is of great importance. Currently, patients with CRC have low survival rates and poor prognoses. Precise categorization of CRC patients based on their prognosis could assist in identifying the most suitable treatment for each individual.

One of the most challenging issues in the clinical field is finding biomarkers with different values for cancer management. In this regard, epigenetic factors that are effective in regulating gene expression are very important [52]. Epigenetic factors refer to factors that can influence gene expression without altering the DNA sequence. Two of the most important epigenetic factors that are effective in regulating gene expression and also important in the field of cancer are miRNAs and DNA methylation. Considering the importance of cell-free nucleic acids and exosomes as the targets of molecular investigations to monitor cancer biomarkers in body fluids, which is also the subject of liquid biopsy, identifying biomarkers with different diagnostic, therapeutic, prognostic, and treatment response

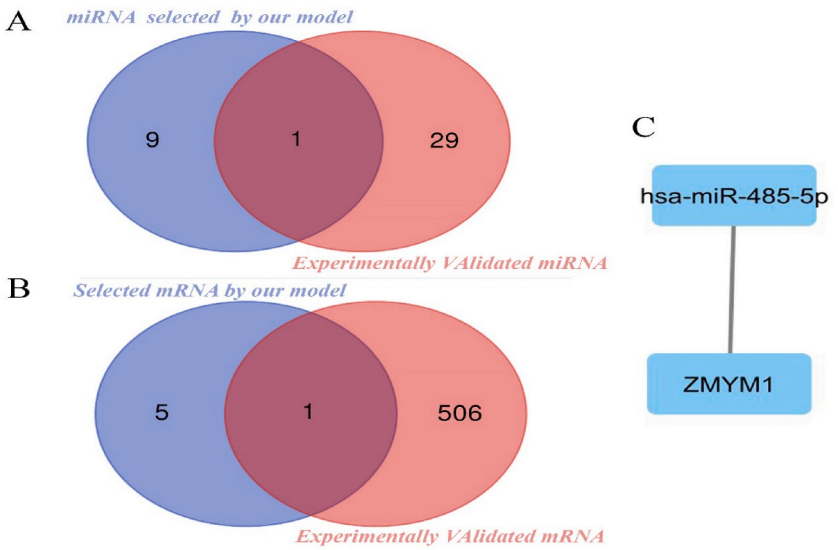


Fig. 8. The relationship between the biomarkers obtained from our study and its investigation in the related database that has been experimentally validated. **A.** The miRNA selected by our study and mirwalk experimentally validated. **B.** The mRNA selected by our study and mirwalk experimentally validated. **C.** the novel connection between the miRNA and mRNA that this study found to be shared and that was experimentally confirmed by other studies, the Mirwalk, and the mirTarbase databases.

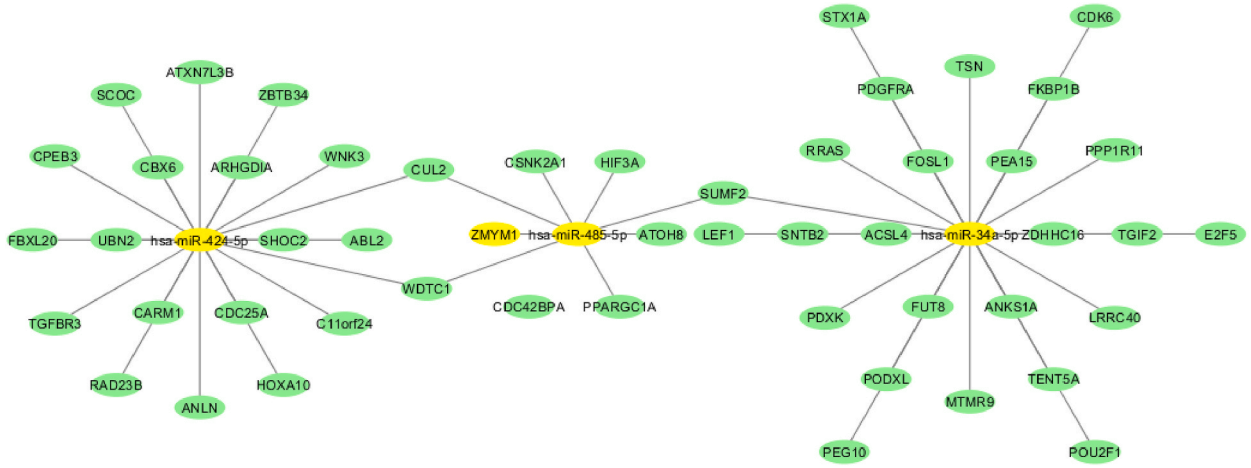


Fig. 9. The bipartite miRNA-mRNA network which the miRNA was the result of our study.

prediction has pivotal importance in cancer management [53,54].

This study proposes a precise colon cancer survival subgroup prediction model using deep-learning techniques and gene expression profiles from human samples. In this case, one of the challenges is the high dimensionality of data which leads to high computational complexity. To address this issue, we used a deep hyperparameter-optimized Autoencoder (HPOAE) to reduce the dimension of input data, and then a rule-based (C5) model was used to identify the survival-associated genes and conditions. The results of this study demonstrate the proposed model's ability to distinguish between high-risk and low-risk groups of colon cancer cases. To evaluate the performance of the model and prevent overfitting, a model was trained by TCGA COAD samples and was confirmed by E-GEOD-17538 as external independent data.

In both datasets, low-risk and high-risk groups were separated clearly. The best deep-learning model, called autoencoder, would return a low MSE value. To achieve this goal, numerous hyperparameters had to be set, and we obtained the best set in comparison to other studies [40,41] by using, and customizing, a hyperparameter optimizer. The findings of this study revealed a higher performance and lower MSE for the proposed model. The HPOAE model has significantly improved the MSE of the model by achieving a minimum MSE of 0.015 as shown in Fig. 3. The lower value of MSE indicates the model has the lowest error rate in the reconstruction of the input layer, and the dense layer is produced with a higher percentage of confidence. By comparing the results of the proposed HPOAE of this study as mentioned in.

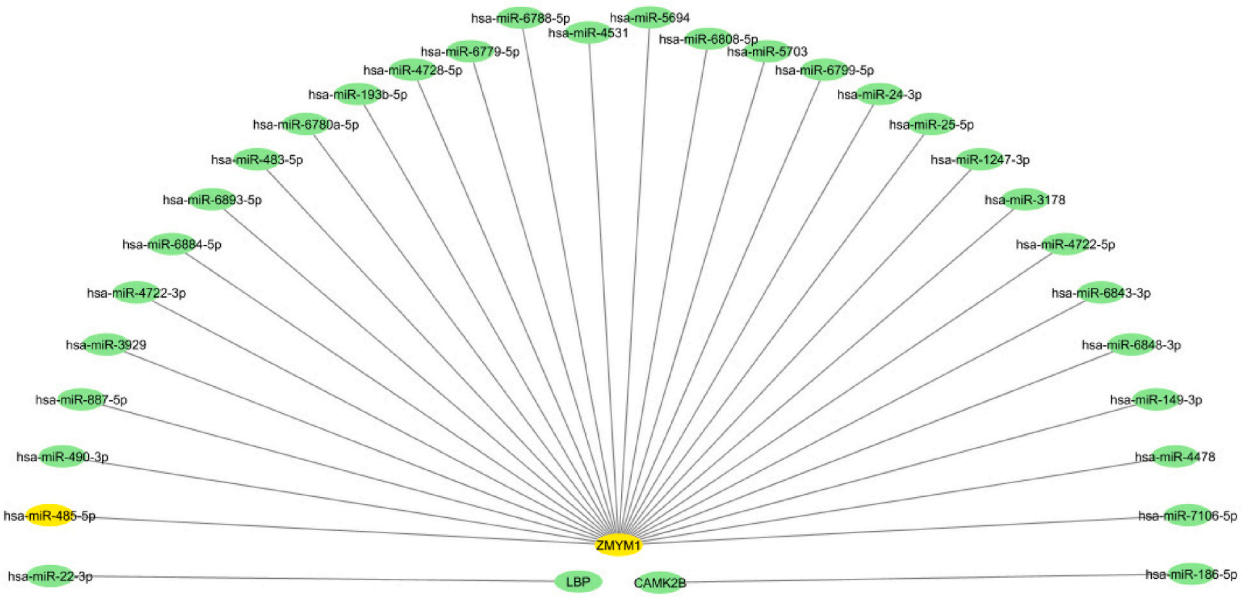


Fig. 10. mRNA-miRNA bipartite network which mRNA was the result of our study.

(Table 5), with those who compared the MSE of the non-Hyper parameter optimized autoencoder model [38] (using a larger sample size of n = 360 compared with the present study with n = 211), a smaller MSE was obtained using a much smaller sample size (n = 211), indicating the need for a noteworthy MSE of the HPOAE models.

Survival-associated features are obtained by segmentation of redimentioned (Dense) TCGA data. Of course, the nature of these data is completely different from the initial data and they are used as projected features in survival analysis. The results of COx-ph analysis separate a sample into high and low-risk groups as mentioned in Figs. 5 and 6 and from this label used for subsequent analyses.

The Supervised Feature selection used the C5 identified a set of 10 miRNAs (including hsa-miR-3130-5p, hsa-miR-3189-5p, hsa-miR-33b-3p, hsa-miR-34a-5p, hsa-miR-365b-5p, hsa-miR-3940-5p, hsa-miR-424-5p, hsa-miR-485-5p, hsa-miR-6842-5p, and hsa-miR-7704), 28 mRNA and 11 methylated genes. By mapping the targets of miRNA to selected genes, six mRNA genes (including ACTL10, CAMK2B, CCL24, FAM122C, LBP, and ZMYM1) were identified that were predicted by other studies, and among these genes, ZMYM1 was validated by experimental assays as well as by the mirTarBase[5] that is thought to be associated with survival in colon cancer samples. Methylation of ZMYM1 (Zinc finger MYM-type containing 1) occurs by METTL3. It is reported that metastasis of the gastric cancer cells increases by binding ZMYM1 to a promoter of E-cadherin [55].

The ZMYM1 is a protein-coding gene type that RefSeq validated. Zinc finger MYM-type protein has protein dimerization activity [56] and ACTL10 is a member of the actin family. Regulation of cell proliferation, migration, and differentiation are some of the roles actin-like proteins serve [56]. The tumor suppressor and the prognostic role of hsa has been reported in cytogenetic normal acute myeloid leukemia [55–57]. Also, a poor prognosis of ACTL10 has been reported in numerous types of cancer, including hepatocellular carcinoma, head and neck squamous cell carcinoma, rhabdomyosarcoma, osteosarcoma, and glioma [56].

Another gene found in our results was CAMK2B. This gene affects colon cancer by regulating proliferation and migration via ERK1/2 and p38 pathways [58]. Furthermore, CAMK2B is a calcium/calmodulin-dependent protein kinase (CaM kinase) II beta-protein [59]. Expression of CCL24 is reported in some tumor cells, such as hepatocellular carcinoma, cutaneous T-cell lymphoma, and colon cancer. Also, It is suggested that this gene can be used as a potential biomarker in colon cancer [60]. CCL24 is a C–C motif chemokine 24 protein and CCR3 chemokine receptor binding. Eotaxin-3 is one of three related chemokines that specifically activate chemokine receptor CCR3 [61]. The protein encoded by FAM122C (Accession GO:0004865), has protein serine/threonine phosphatase inhibitor activity, binds to and stops, prevents, or reduces the activity of a serine/threonine protein phosphatase, an enzyme that catalyzes the reaction of protein serine/threonine phosphate and maintains the growth of hepatocellular carcinoma cells through promoting MAPK/AKT signaling [62].

LBP (GO:0071723) is a lipopeptide binding, binding to a lipopeptide any of a group of organic compounds comprised of two or more amino acids linked by peptide bonds and containing a no-protein group consisting of lipid or lipids [62]. The lipopolysaccharide (LPS)-binding protein (LBP) gene plays a crucial role in the innate immune response to the development of inflammatory and infectious-related diseases. LBP expression was positively associated with tumor node metastasis stage and tumor differentiation [63].

Based on the study of methylated genes, the following methylated genes are in the group of genes targeted by the predicted target

[5] https://mirtarbase.cuhk.edu.cn/.

scan, mirDB.[6] The important point is that the ANGPT4 gene is in the group of target genes validated by mirTarBase, and in the results in the DiseaseMeth[7] database.

Among the identified miRNAs, every two miRNAs targeting at least one common gene, for example, hsa-miR-33b-5p, and hsa-miR-3940-5p, have a common target, AGO2. The genes, hsa-miR-485-5p and hsa-miR-7704 are targeting APOL6, hsa-miR-424-5p, hsa-miR-485-5p, and have a common target, ARHGDIA. Other miRNA targets are shown in (Table 5). The miR-33b gene is a member of the miR-33 family and it acts as a tumor suppressor in different cancers such as TNBC, non-small-cell lung cancer, colorectal cancer, and esophageal squamous cell carcinoma via targeting EMT and proliferation [64]. Poor expression of miR-3940-5p is reported in CRC patient serum and is closely linked to the prognosis for patients [65]. The gene, miR-485-5p, provides a tumor-suppressing microRNA in some human cancers. Overexpression of miR-485-5p could inhibit the proliferation and invasion of CRC cell lines in vitro and in vivo. Also, it is reported that miR-485-5p plays a pivotal tumor-suppressing role in CRC [66]. Up-regulation of hsa-miR-3189 has been reported in colon cancer. Some experimental studies have demonstrated that the change in expression of hsa-mir-3189 has played crucial roles in other cancer cells [67]. It is reported that better survival of patients is associated with high miR-34a expression in colorectal cancer [68]. By targeting AKT3 and PSAT1, miR-424 has a tumor-suppressive role in human colorectal cancer [69]. By targeting CD147, hsa-miR-485-5p has a tumor suppression function in colorectal cancer cells [66]. Based on previously mentioned data and studies, we can suggest that these miRNAs and genes can be used as diagnostic biomarkers of high-risk and low-risk patients with colon cancer. No other information about FAM122C, hsa-miR-365b-5p, and hsa-miR-6842-5p in cancer-related studies has been found, therefore, it is suggested that the present findings must be investigated in clinical and experimental studies.

The target prediction score of two different web application prediction tools, Targetscan[8], and MR-microT[9], which are well known, was used for predicting the target genes of miR-6842 and miR-33b-3p. It indicates that ACTL10 and LBP have acceptable validation scores, although these targets are not validated by the mirTarBase[10] database. They can be good primary targets for experimental studies. We studied various expression data such as miRNA, mRNA, and methylate in patients with colon cancer. The interesting point was that tp53 is one of the targets of hsa-miR-485-5p, and hsa-miR-485-5p targets exactly the selected ZMYM1 gene, which is predicted by our model and validated by mirTarBase.

In the present study, changes in the expression of 6 genes are known to be associated with the survival of patients with colon cancer.

The interesting point is that all these genes are involved in the process of cancer malignancy and metastasis. The mechanism of the effect of these genes is mainly through the stimulation of EMT and directing the path toward metastasis [53,70,71].

Among other interesting topics with clinical value, which is the result of the present study, we can mention the epigenetic factors effective in regulating the expression of the genes that are important in cancer development such as miR-3940-5p and miR-485-5p which is a good prognostic factor indicating better survival in colon cancer.

On the other hand, promoter hypermethylation of genes that are the target of candidate miRNAs in this study confirms that the epigenetic regulations of these genes are very important in colon cancer.

The model's performance is impacted by the heterogeneity of the population. However, there are challenges in testing these models on external datasets since the submodels were constructed using small training data and may suffer from over-fitting in confirmation cohorts. Additionally, sample risk factors may not always be available for public cohorts, limiting our ability to confirm the model's performance. Despite these challenges, the TCGA-based HPOAE model has a Log Rank Mantel-Cox value of 14.27 ± 2, indicating its overall predictive capability. We also used other performance metrics such as a Breslow-Generalized Wilcoxon value of 13.13 ± 1 to evaluate our pipeline. In the future, we plan to collaborate with clinicians to develop prospective cohorts and continually enhance the model.

One important limitation of this study was to not considering clinical features of the patients as well as imaging data on the cancer tissue in the optimizd autoencoder and only included DNA methylation, RNA-seq, and miRNA-seq data related to only $n = 368$ sample. Larger samples may improve the results. One possible extension of the present study is to add other modalities such as CT-scan that provides image data as another data type to the network and integrate this type of data along with omics data to provide a better insight into patients' survival prediction based on much larger sample sizes. Also, considering supervised deep learning methods and comparing them with the proposed approach is an interesting research area for future research topics.

Overall, this study is an important contribution to the field of cancer research, and its findings may provide a valuable foundation for further investigation into the development of effective treatments for colon cancer.

This study presents a comprehensive investigation utilizing published data and bioinformatic analysis. In vitro or in vivo models will be necessary to validate the findings. The results of this study are expected to facilitate future research efforts.

6. Conclusion

In the present study, a deep-learning method was used, and by optimizing the model's parameters, the MSE rate was improved compared to conventional and similar models. Using these results, high-risk and low-risk groups can be identified. In addition, colon cancer-associated biomarkers can be used separately for methylation, mRNA, and miRNAs for early detection of this cancer. These

[6] http://mirdb.org/.
[7] http://bio-bigdata.hrbmu.edu.cn.
[8] http://www.targetscan.org.
[9] http://diana.imis.athena-innovation.gr.
[10] https://mirtarbase.cuhk.edu.cn/.

findings can lead to improved outcomes for CRC patients. In the future, by improving computing power and new studies based on deep learning and different databases, more suitable methods can be developed to increase productivity and improve efficiency for the survival of people with colon cancer. Moreover, future clinical and experimental studies could use the results of this study and its methods. Finally, we can now predict cancer in its early stages, and the survival of patients and their treatment can be improved.

Data and code availability

We use TCGAbiolinks [72] to download colon adenocarcinoma (COAD) from *The Cancer Genome Atlas* (*TCGA*). Google collab (colab.research.google.com/) to implement deep learning model by Jupiter and python. The STRING database is accessible at https://string-db.org. The MIRWALK database is accessible at http://mirwalk.umm.uni-heidelberg.de/. The mirtarbase database is accessible at https://mirtarbase.cuhk.edu.cn/. The TARGET SCAN database is accessible at http://www.targetscan.org.The MR-microT database is accessible at http://diana.imis.athena-innovation.gr.

Author contributions

A.MN and A.N: Conceived and designed the study.
S.S: Performed the experiments.
S.S, K.A and H.L: Analyzed and interpreted the data.
S.S and H.L: Contributed analysis tools, or data.
S.S, A.MN, M.S and L.T: Drafted the article and critically revised its important intellectual content.
All authors have reviewed and approved the final version of the manuscript for submission.

Declaration of competing interest

None Declared

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e17653.

References

[1] Chemokines in colon cancer progression, in: S.-N. Jia, Y.-B. Han, R. Yang, Z.-C. Yang (Eds.), Seminars in Cancer Biology, Elsevier, 2022.
[2] C.M. Johnson, C. Wei, J.E. Ensor, D.J. Smolenski, C.I. Amos, B. Levin, et al., Meta-analyses of colorectal cancer risk factors, Cancer Causes Control 24 (6) (2013) 1207–1222.
[3] N. Howlader, M. Krapcho, D. Miller, K. Bishop, C. Kosary, M. Yu, K.A. Cronin (Eds.), SEER Cancer Statistics Review, 1975-2014, National Cancer Institute, Bethesda, MD, 2017.
[4] A.B. Benson, A.P. Venook, M.M. Al-Hawary, M.A. Arain, Y.-J. Chen, K.K. Ciombor, et al., Colon cancer, version 2.2021, NCCN clinical practice guidelines in oncology, J. Natl. Compr. Cancer Netw. 19 (3) (2021) 329–359.
[5] R. Deventhiran, S. Senthilkumar, EXPLORING COLORECTAL CANCER GENES THROUGH DATA MINING TECHNIQUES, 2019.
[6] S. Bach, I. Paulis, N. Sluiter, M. Tibbesma, I. Martin, M. van de Wiel, et al., Detection of colorectal cancer in urine using DNA methylation analysis, Sci. Rep. 11 (1) (2021) 1–11.
[7] E. Van Cutsem, A. Cervantes, R. Adam, A. Sobrero, J. Van Krieken, D. Aderka, et al., ESMO consensus guidelines for the management of patients with metastatic colorectal cancer, Ann. Oncol. 27 (8) (2016) 1386–1422.
[8] D.P. Harrington, T.R. Fleming, A class of rank test procedures for censored survival-data, Biometrics 37 (1981) 613.
[9] T.A. Manolio, Genomewide association studies and assessment of the risk of disease, N. Engl. J. Med. 363 (2) (2010) 166–176.
[10] S.J. Winawer, A.G. Zauber, The advanced adenoma as the primary target of screening, Gastrointest. Endosc. Clin. 12 (1) (2002) 1–9.
[11] R. Gor, S.S. Sampath, L.M. Lazer, S. Ramalingam, RNA binding protein PUM1 promotes colon cancer cell proliferation and migration, Int. J. Biol. Macromol. 174 (2021) 549–561.
[12] S. Liu, S. Gao, T. Liu, J. Yang, X. Zheng, Z. Li, Circular RNA SMARCA5 functions as an anti-tumor candidate in colon cancer by sponging microRNA-552, Cell Cycle (2021) 1–13.
[13] S. Jin, D. Zhu, F. Shao, S. Chen, Y. Guo, K. Li, et al., Efficient detection and post-surgical monitoring of colon cancer with a multi-marker DNA methylation liquid biopsy, Proc. Natl. Acad. Sci. USA 118 (5) (2021).
[14] H.S. Yoo, S.B. Won, Y.H. Kwon, Luteolin induces apoptosis and autophagy in HCT116 colon cancer cells via p53-dependent pathway, Nutr. Cancer (2021) 1–10.
[15] Y. Masoudi-Sobhanzadeh, Y. Omidi, M. Amanlou, A. Masoudi-Nejad, DrugR+: a comprehensive relational database for drug repurposing, combination therapy, and replacement therapy, Comput. Biol. Med. 109 (2019) 254–262.
[16] I. Subramanian, S. Verma, S. Kumar, A. Jere, K. Anamika, Multi-omics data integration, interpretation, and its application, Bioinf. Biol. Insights 14 (2020), 1177932219899051.
[17] S. Huang, K. Chaudhary, L.X. Garmire, More is better: recent progress in multi-omics data integration methods, Front. Genet. 8 (2017) 84.
[18] M. Kouhsar, S. Azimzadeh Jamalkandi, A. Moeini, A. Masoudi-Nejad, Detection of novel biomarkers for early detection of Non-Muscle-Invasive Bladder Cancer using Competing Endogenous RNA network analysis, Sci. Rep. 9 (1) (2019) 1–15.
[19] H. Motieghader, M. Kouhsar, A. Najafi, B. Sadeghi, A. Masoudi-Nejad, mRNA–miRNA bipartite network reconstruction to predict prognostic module biomarkers in colorectal cancer stage differentiation, Mol. Biosyst. 13 (10) (2017) 2168–2180.
[20] H. Torkey, M. Atlam, N. El-Fishawy, H. Salem, A novel deep autoencoder based survival analysis approach for microarray dataset, PeerJ Computer Science 7 (2021) e492.
[21] R. Zemouri, N. Zerhouni, D. Racoceanu, Deep learning in the biomedical applications: recent and future status, Appl. Sci. 9 (8) (2019) 1526.
[22] K. Cao, Y. Hong, L. Wan, Manifold alignment for heterogeneous single-cell multi-omics data integration using Pamona, Bioinformatics 38 (1) (2022) 211–219.

[23] M. Kang, E. Ko, T.B. Mersha, A roadmap for multi-omics data integration using deep learning, Briefings Bioinf. 23 (1) (2022) bbab454.
[24] S.Y. Kim, H.-H. Jeong, J. Kim, J.-H. Moon, K.-A. Sohn, Robust pathway-based multi-omics data integration using directed random walks for survival prediction in multiple cancer studies, Biol. Direct 14 (1) (2019) 1–13.
[25] V. Malik, Y. Kalakoti, D. Sundar, Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer, BMC Genom. 22 (1) (2021) 1–11.
[26] L. Tong, H. Wu, M.D. Wang, Integrating multi-omics data by learning modality invariant representations for improved prediction of overall survival of cancer, Methods 189 (2021) 74–85.
[27] Talia Golan, et al., Genomic features and classification of homologous recombination deficient pancreatic ductal adenocarcinoma, Gastroenterology 160 (6) (2021) 2119–2132.
[28] A.W. Smith, M. Gallitto, E.J. Lehrer, I. Wasserman, V. Gupta, S. Sharma, et al., Redefining risk of contralateral cervical nodal disease in early stage oropharyngeal cancer in the human papillomavirus era, Head Neck 43 (5) (2021) 1409–1414.
[29] H. Ueno, M. Ishiguro, E. Nakatani, T. Ishikawa, H. Uetake, K. Murotani, et al., Prognostic value of desmoplastic reaction characterisation in stage II colon cancer: prospective validation in a Phase 3 study (SACURA Trial), Br. J. Cancer 124 (6) (2021) 1088–1097.
[30] C.-F.J. Yang, A. Kumar, J.Z. Deng, V. Raman, N.S. Lui, T.A. D'Amico, et al., A national analysis of short-term outcomes and long-term survival following thoracoscopic versus open lobectomy for clinical stage II non-small-cell lung cancer, Ann. Surg. 273 (3) (2021) 595–605.
[31] S. Ammendola, G. Turri, I. Marconi, G. Burato, S. Pecori, A. Tomezzoli, et al., The presence of poorly differentiated clusters predicts survival in stage II colorectal cancer, Virchows Arch. 478 (2) (2021) 241–248.
[32] T. Gan, K.B. Schaberg, D. He, A. Mansour, H. Kapoor, C. Wang, et al., Association between obesity and histological tumor budding in patients with nonmetastatic colon cancer, JAMA Netw. Open 4 (4) (2021), e213897-e.
[33] M. Ghasemi, H. Seidkhani, F. Tamimi, M. Rahgozar, A. Masoudi-Nejad, Centrality measures in biological networks, Curr. Bioinf. 9 (4) (2014) 426–441.
[34] P. Razzaghi, K. Abbasi, M. Shirazi, S. Rashidi, Multimodal brain tumor detection using multimodal deep transfer learning, Appl. Soft Comput. 129 (2022), 109631.
[35] P. Razzaghi, K. Abbasi, J.B. Ghasemi, Multivariate Pattern Recognition by Machine Learning Methods. Machine Learning and Pattern Recognition Methods in Chemistry from Multivariate and Data Driven Modeling, Elsevier, 2023, pp. 47–72.
[36] Deep autoencoder neural networks for gene ontology annotation predictions, in: D. Chicco, P. Sadowski, P. Baldi (Eds.), Proceedings of the 5th ACM Conference on Bioinformatics, computational biology, and health informatics, 2014.
[37] C.S.N. Pathirage, J. Li, L. Li, H. Hao, W. Liu, P. Ni, Structural damage identification based on autoencoder neural networks and deep learning, Eng. Struct. 172 (2018) 13–28.
[38] G. Melis, C. Dyer, P. Blunsom, On the State of the Art of Evaluation in Neural Language Models, 2017 arXiv preprint arXiv:170705589.
[39] J. Snoek, H. Larochelle, R.P. Adams, Practical bayesian optimization of machine learning algorithms, Adv. Neural Inf. Process. Syst. 25 (2012).
[40] K. Chaudhary, O.B. Poirion, L. Lu, L.X. Garmire, Deep learning–based multi-omics integration robustly predicts survival in liver cancer, Clin. Cancer Res. 24 (6) (2018) 1248–1259.
[41] J. Lv, J. Wang, X. Shang, F. Liu, S. Guo, Survival prediction in patients with colon adenocarcinoma via multiomics data integration using a deep learning algorithm, Biosci. Rep. 40 (12) (2020).
[42] P. Kokoszka, R. Kulik, Principal component analysis of infinite variance functional data, J. Multivariate Anal. 193 (2023), 105123.
[43] M.P. Libório, O. da Silva Martinuci, A.M.C. Machado, T.M. Machado-Coelho, S. Laudares, P. Bernardes, Principal component analysis applied to multidimensional social indicators longitudinal studies: limitations and possibilities, Geojournal 87 (3) (2022) 1453–1468.
[44] N. Sapoval, A. Aghazadeh, M.G. Nute, D.A. Antunes, A. Balaji, R. Baraniuk, et al., Current progress and open challenges for applying deep learning across the biosciences, Nat. Commun. 13 (1) (2022) 1728.
[45] A. Singh, T. Ogunfunmi, An overview of variational autoencoders for source separation, finance, and bio-signal applications, Entropy 24 (1) (2022) 55.
[46] Y. Bengio, Learning Deep Architectures for AI, Now Publishers Inc, 2009.
[47] G. Louppe, Bayesian Optimisation with Scikit-Optimize, 2017.
[48] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.
[49] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[50] Y. Masoudi-Sobhanzadeh, H. Motieghader, A. Masoudi-Nejad, FeatureSelect: a software for feature selection based on machine learning approaches, BMC Bioinf. 20 (1) (2019) 170.
[51] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.
[52] G. Guo, Z. Tan, Y. Liu, F. Shi, J. She, The therapeutic potential of stem cell-derived exosomes in the ulcerative colitis and colorectal cancer, Stem Cell Res. Ther. 13 (1) (2022) 1–18.
[53] O.Y. Gorlova, M. Kimmel, S. Tsavachidis, C.I. Amos, I.P. Gorlov, Identification of lung cancer drivers by comparison of the observed and the expected numbers of missense and nonsense mutations in individual human genes, Oncotarget 13 (2022) 756.
[54] J.F. Chen, Q. Yan, The roles of epigenetics in cancer progression and metastasis, Biochem. J. 478 (17) (2021) 3373–3393.
[55] S. Xie, W. Chen, K. Chen, Y. Chang, F. Yang, A. Lin, et al., Emerging roles of RNA methylation in gastrointestinal cancers, Cancer Cell Int. 20 (1) (2020) 1–11.
[56] R. Lai, W. Zhang, X. He, X. Liao, X. Liu, W. Fu, et al., Prognostic role of ACTL10 in cytogenetic normal acute myeloid leukemia, J. Cancer 11 (17) (2020) 5150.
[57] L. Chen, Y.-H. Zhang, Z. Zhang, T. Huang, Y.-D. Cai, Inferring novel tumor suppressor genes with a protein-protein interaction network and network diffusion algorithms, Molecular Therapy-Methods & Clinical Development 10 (2018) 57–67.
[58] W. Chen, P. An, X.-J. Quan, J. Zhang, Z.-Y. Zhou, L.-P. Zou, et al., Ca2+/calmodulin-dependent protein kinase II regulates colon cancer proliferation and migration via ERK1/2 and p38 pathways, World J. Gastroenterol. 23 (33) (2017) 6111.
[59] M. Doroudi, M.C. Plaisance, B.D. Boyan, Z. Schwartz, Membrane actions of 1α, 25 (OH) 2D3 are mediated by Ca2+/calmodulin-dependent protein kinase II in bone and cartilage cells, J. Steroid Biochem. Mol. Biol. 145 (2015) 65–74.
[60] S.-J. Lim, CCL24 Signaling in the Tumor Microenvironment. Tumor Microenvironment, Springer, 2021, pp. 91–98.
[61] J. Ye, K.L. Mayer, M.R. Mayer, M.J. Stone, NMR solution structure and backbone dynamics of the CC chemokine eotaxin-3, Biochemistry 40 (26) (2001) 7820–7831.
[62] K. Takeda, T. Kaisho, S. Akira, Toll-like receptors, Annu. Rev. Immunol. 21 (2003) 335–376.
[63] Q.Y. Cai, J.H. Jiang, R.M. Jin, G.Z. Jin, N.Y. Jia, The clinical significance of lipopolysaccharide binding protein in hepatocellular carcinoma, Oncol. Lett. 19 (1) (2020) 159–166.
[64] B. Pattanayak, I. Garrido-Cano, A. Adam-Artigues, E. Tormo, B. Pineda, P. Cabello, et al., MicroRNA-33b suppresses epithelial–mesenchymal transition repressing the MYC–EZH2 pathway in HER2+ breast carcinoma, Front. Oncol. (2020) 1661.
[65] T. Li, Y. Wan, Z. Su, J. Li, M. Han, C. Zhou, Mesenchymal stem cell-derived exosomal microRNA-3940-5p inhibits colorectal cancer metastasis by targeting integrin α6, Dig. Dis. Sci. 66 (6) (2021) 1916–1927.
[66] Y. Pan, H. Sun, X. Hu, B. He, X. Liu, T. Xu, et al., The inhibitory role of miR-485-5p in colorectal cancer proliferation and invasion via targeting of CD147, Oncol. Rep. 39 (5) (2018) 2201–2208.
[67] W. Chen, C. Gao, Y. Liu, Y. Wen, X. Hong, Z. Huang, Bioinformatics analysis of prognostic miRNA signature and potential critical genes in colon cancer, Front. Genet. 11 (2020) 478.
[68] K. Hasakova, R. Reis, M. Vician, M. Zeman, I. Herichova, Expression of miR-34a-5p is up-regulated in human colorectal cancer and correlates with survival and clock gene PER2 expression, PLoS One 14 (10) (2019), e0224396.
[69] Y. Fang, X. Liang, J. Xu, X. Cai, miR-424 targets AKT3 and PSAT1 and has a tumor-suppressive role in human colorectal cancer, Cancer Manag. Res. 10 (2018) 6537.

[70] Q. He, Z. Li, The dysregulated expression and functional effect of CaMK2 in cancer, Cancer Cell Int. 21 (2021) 1–15.
[71] S.-J. Lim, CCL24 signaling in the tumor microenvironment, Tumor Microenvironment: The Role of Chemokines–Part B (2021) 91–98.
[72] A. Colaprico, T.C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, et al., TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data, Nucleic Acids Res. 44 (8) (2016) e71–e.