



OPEN

Machine learning algorithms reveal potential miRNAs biomarkers in gastric cancer

Hanieh Azari^{1,3,10}, Elham Nazari^{1,2,10}, Reza Mohit⁴, Alireza Asadnia^{1,3}, Mina Maftooh¹, Mohammadreza Nassiri⁵, Seyed Mahdi Hassanian^{1,2}, Majid Ghayour-Mobarhan¹, Soodabeh Shahidsales⁶, Majid Khazaei^{1,2,10}, Gordon A. Ferns^{7,8}✉ & Amir Avan^{1,8,9}✉

Gastric cancer is the high mortality rate cancers globally, and the current survival rate is 30% even with the use of combination therapies. Recently, mounting evidence indicates the potential role of miRNAs in the diagnosis and assessing the prognosis of cancers. In the state-of-art research in cancer, machine-learning (ML) has gained increasing attention to find clinically useful biomarkers. The present study aimed to identify potential diagnostic and prognostic miRNAs in GC with the application of ML. Using the TCGA database and ML algorithms such as Support Vector Machine (SVM), Random Forest, k-NN, etc., a panel of 29 was obtained. Among the ML algorithms, SVM was chosen (AUC:88.5%, Accuracy:93% in GC). To find common molecular mechanisms of the miRNAs, their common gene targets were predicted using online databases such as miRWalk, miRDB, and Targetscan. Functional and enrichment analyzes were performed using Gene Ontology (GO) and Kyoto Database of Genes and Genomes (KEGG), as well as identification of protein–protein interactions (PPI) using the STRING database. Pathway analysis of the target genes revealed the involvement of several cancer-related pathways including miRNA mediated inhibition of translation, regulation of gene expression by genetic imprinting, and the Wnt signaling pathway. Survival and ROC curve analysis showed that the expression levels of hsa-miR-21, hsa-miR-133a, hsa-miR-146b, and hsa-miR-29c were associated with higher mortality and potentially earlier detection of GC patients. A panel of dysregulated miRNAs that may serve as reliable biomarkers for gastric cancer were identified using machine learning, which represents a powerful tool in biomarker identification.

Gastric cancer (GC) has a low 5-year survival rate in part due to misdiagnosis in the early stages of the disease and its recurrence, which makes this cancer one of the most prevalent cancers and the leading cause of cancer-related mortality¹. Misdiagnosis of the GC in the early stages, when no symptoms are apparent, and only the accumulation of genetic changes occurs at the molecular level, brings a heavy financial and physical burden to patients thus emphasizing the need for a clear understanding of the underlying molecular pathways². Among the components of this molecular network are the miRNAs, "short-length post-transcriptional regulators," which are part of the dynamic genome changes that act in the initiation and progression of cancers such as GC^{3,4}. Recent advances in genome-wide sequencing and the use of bioinformatics tools to analyze their data have led to new roads to further identifying components and their functions in this intertwined molecular network. Due to the complex nature of the disease, the current biomarkers lack the necessary power for accurate diagnosis and prognosis, as a result, more studies have been directed towards the discovery of computational methods such as machine learning to predict the miRNA-disease relationship and their use in cancer diagnosis⁵. In a recent study, the use of Boruta's machine learning algorithm in a GEO dataset of gastric cancer patients introduced 30

¹Metabolic Syndrome Research Center, Mashhad University of Medical Sciences, Mashhad, Iran. ²Basic Sciences Research Institute, Mashhad University of Medical Sciences, Mashhad, Iran. ³Medical Genetics Research Center, Mashhad University of Medical Sciences, Mashhad, Iran. ⁴Department of Anesthesia, Bushehr University of Medical Sciences, Bushehr, Iran. ⁵Recombinant Proteins Research Group, The Research Institute of Biotechnology, Ferdowsi University of Mashhad, Mashhad, Iran. ⁶Cancer Research Center, Mashhad University of Medical Sciences, Mashhad, Iran. ⁷Division of Medical Education, Brighton and Sussex Medical School, Falmer, Brighton, Sussex BN1 9PH, UK. ⁸Faculty of Health, School of Biomedical Sciences, Queensland University of Technology, Brisbane, Australia. ⁹College of Medicine, University of Warith Al-Anbiyaa, Karbala, Iraq, College of Medicine, University of Warith Al-Anbiyaa, karbala, Iraq. ¹⁰These authors contributed equally: Hanieh Azari, Elham Nazari and Majid Khazaei. ✉email: G.Ferns@bsms.ac.uk; avana@mums.ac.ir

miRNAs with high biomarker power, and finally, hsa-miR-1343-3p has been introduced as a valuable diagnostic biomarker⁶. In another study, Rehman et al. investigated clinically validated miRNAs, using machine learning. The results not only confirmed their concept but also indicated the potential of miRNA as diagnostic biomarkers in breast cancer and introducing machine learning as a powerful tool in functional studies⁷. Utilizing data from the study of human tumors in large cohorts, the TCGA project has developed an extensive catalog of genetic changes associated with these tumors, make cancer genome and transcriptome uncovered and shed new light on the diagnosis, treatment, and prognosis of cancer, as in gastric cancer with the help of accurate classification of gastric adenoma cancers into four subgroups, (EBV-positive tumors (EBV), microsatellite unstable tumors (MSI), genomically stable tumors and chromosomally unstable tumors (CIN) has led to improved clinical diagnosis and treatment in this cancer⁸. This study aimed to find differentially expressed miRNAs (demiRs) in gastric cancer which can act as an important biomarker for early detection.

Material and methods

Retrieval of genome data. Raw data from clinical information and sequencing of 576 stomach cancer were downloaded from the TCGA database (<https://tcga-data.nci.nih.gov/>), an available dataset that catalogs cancer-causing genomic changes⁸.

Data preprocessing and identification of DemiRs. Processing and analyzing of raw data were accomplished using R software (version 4.2.0, Limma package, <https://www.r-project.org/>), to identify differentially expressed miRNAs (demiRs), adj. $p < 0.05$ and $\log|FC| > 1.5$ were set as the threshold.

Identifying prognostic biomarkers. The association between the expression level of previously identified demiRs and the overall survival rate and outcome of the patients was assessed by log-rank analysis ($p < 0.05$). To define and visualize miRNAs with prognostic ability, the Kaplan Meier was used. Survival, survminer, and ggplot2 R packages were used for this purpose.

Identifying predictive biomarkers. Machine learning is defined as the intelligent analysis and design of models with the ability to find features with significant impact and discover the relationships between them based on algorithms and mass data analysis, which has comparable capabilities with human experts⁹. One of its several applications is in medical sciences to identify biomarkers with significant diagnostic value for various diseases as in the wide range of cancers such as colorectal cancer, breast cancer, gastric cancer, as well as cardiovascular disease, Alzheimer's disease, and etc.^{5,10–13}. In this study, we used machine learning to find miRNAs with diagnostic and predictive values. For this purpose, in the processing step, important features were identified using heatmap analysis, and five methods (SVM, DTS, RF, Logistic Regression, KNN) were used for classification; in the next step, all of these methods were compared with four different metrics (accuracy, f1score, ROC_curve, and confusion matrix) to find the most accurate algorithm. The algorithms were used in this study to measure the diagnostic ability of demiRs in GC which will be briefly mentioned in the following. One of the methods of evaluating the performance of binary classification is the "Receiver Operating Characteristic" or ROC curve. The efficiency of "Binary Classifier" algorithms is usually measured by indicators called "Sensitivity" or "Recall" but in the ROC chart, both of these indicators are combined and displayed as a curve. In the current study, the ROC curve is used to evaluate the efficiency diagnostic of selected hub miRNAs using R 4.2.2's combiROC package. For this purpose, sensitivity, specificity, cut-off value, positive predictive value, negative predictive value, and area under the ROC curve were measured. A support vector machine (SVM) algorithm is used for the classification of data points as well as regression in machine learning. SVMs have played an important role in a wide variety of biological applications, for example, the classification of microarray gene expression profiles in tumor and normal samples to find suitable diagnostic or prognostic biomarkers¹⁴. The K-Nearest Neighbors (KNN) algorithm is one of the simplest and at the same time the best algorithms used in supervised learning in the field of machine learning which considers the distance in classification and uses both "Regression" and "Classification" issues¹⁵. One of the methods of classification in "Supervised Machine Learning" is logistics regression. This method of regression, relies on an odd ratio for calculation. This type of analysis can help predict the probability of occurring an event or choice¹⁶. DTs: A decision tree is a tool for making a more appropriate decision in a way that gives a tree structure or hierarchical structure to decisions and their results. The structure of this tree can also be based on chance and probability, so choosing any decision randomly can bring risks or benefits¹⁷. Random Forest is an easy-to-use machine learning algorithm that often provides excellent results even without adjusting its meta-parameters. This algorithm is considered one of the most used machine learning algorithms due to its simplicity and usability, both for "Classification" and "Regression"¹⁴.

Machine learning performances. The accuracy parameter is the most commonly used, basic, and simple measure of the quality of a category, and it is the amount of correct recognition of the category in a total of two categories¹⁸. "System performance characteristic curve" (Receiver Operating Characteristic | ROC) is an appropriate evaluation index to ensure the validity of the results based on sensitivity and specificity. (Area Under Curve) is used as a measure to evaluate the performance of the category¹⁹. When the accuracy of the diagnosis of a category is more important compared to the accuracy of the overall diagnosis, the concept of "Confusion Matrix" comes to our aid. confusion matrix can give you a better idea of what your classification model is getting right and what types of errors it is making²⁰.

Validation of candidate microRNAs in datasets. In order to validate the candidate miRNAs which resulted by machine learning algorithms, we validated them using online web server such as, Global Data Assembly Centres (GDAC) (<https://gdac.broadinstitute.org/>) and dbDMEC (<https://www.biosino.org/dbDMEC/index>) webtools which contains the Differentially Expressed MiRNAs in human Cancers based on public repositories including Gene Expression Omnibus (GEO), Sequence Read Archive (SRA), ArrayExpress and The Cancer Genome Atlas (TCGA).

miRNA-target prediction. Using online databases for simple prediction of miRNA-targets such as miRwalk (<http://mirwalk.umm.uni-heidelberg.de/>)²¹, miRDB (<http://www.mirdb.org/>)²², and Target scan (https://www.targetscan.org/vert_80/)²³, target genes of significant validated up and down expressed miRNAs were predicted.

PPI network analysis. Search Tool for the Retrieval of Interacting Genes (STRING; <https://string-db.org/>)²⁴ with confidence scores ≥ 0.9 as a threshold was used to find an interaction between the identified miRNAs' targets.

Functional and pathway enrichment analysis. Using functional annotation tools such as (GO; <http://geneontology.org/>)²⁵ which annotate genes considering three aspects biological functions (BF), molecular functions (MF), and cellular components (CC), and Kyoto Encyclopedia of Genes and Genomes (KEGG; <https://www.genome.jp/kegg/>)²⁶, we clarified the functional and enrichment analysis of the candidate genes.

Correlation analysis. Correlation matrix analyses were then performed between miRNAs and clinicopathological characteristics including, gender, age, stage, and prior malignancy using Spearman's correlation method.

Construction of miRNA–mRNA network. To obtain a clear visualization of the miRNAs–mRNAs interaction network, Cytoscape software (version 3.9.1) (<https://cytoscape.org/>)²⁷ was used. Highly interconnected miRNAs and genes were then demonstrated. The flowchart of the methodology used in the current study is demonstrated in Fig. 1.

Results

Demographic information. The clinicopathological information of 348 (64.9%) men and 188 (35.1%) women included in this study were downloaded from TCGA and summarized in Table 1. The average age was 65.3 years and around 250 (46.6%) had advanced gastric cancer.

Correlation analysis. Of the clinicopathological data, only the stage of disease was significantly associated with cancer. The significant criterion for measuring association was considered a p -value of < 0.05 (Fig. 2A).

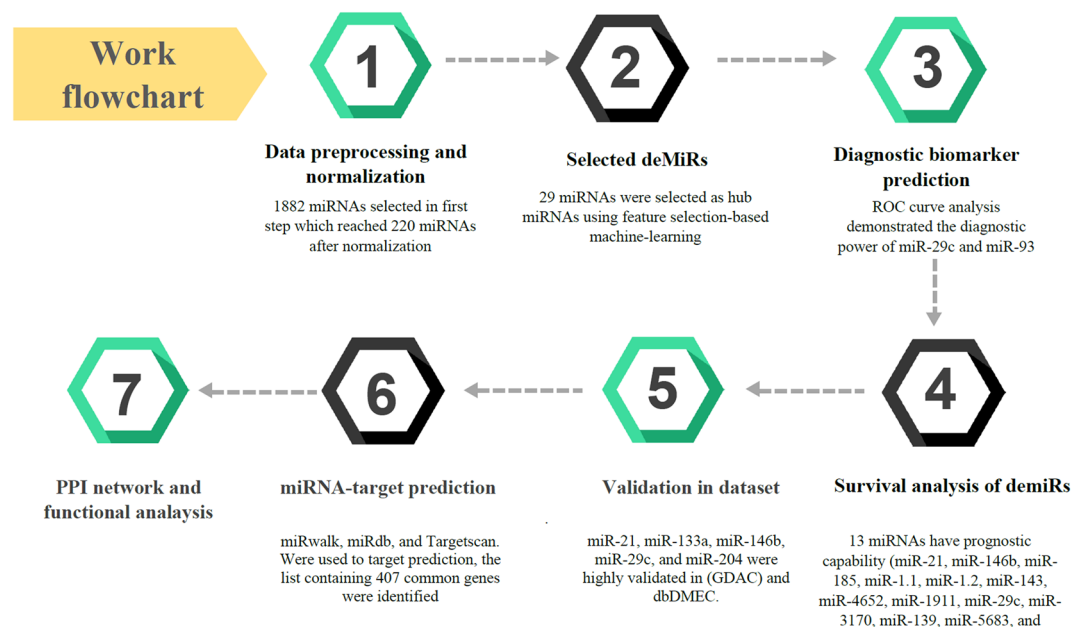


Figure 1. Flowchart of methodology performed in this research.

| Demographic variables | No. of patients (%) / Mean \pm SD |
|---------------------------------|-------------------------------------|
| Patients | 464 |
| Mean age (years, mean \pm SD) | 65.33 \pm 12.091 |
| Gender | |
| Female | 188 (35.1) |
| Male | 348 (64.9) |
| Stage | |
| 1 | 78 (14.6) |
| 2 | 180 (33.6) |
| 3 | 198 (36.9) |
| 4 | 52 (9.7) |
| Prior malignancy | |
| Yes | 13 (2.4) |
| No | 523 (97.6) |

Table 1. Demographic information.

Data collection. As mentioned in the material & method section, the source of the Raw data from clinical information and sequencing was the TCGA database. Based on the mentioned criteria, 536 samples were selected for further studies of which about 465 were related to patients with GC and 72 were related to age and sex-matched control.

Data preprocessing and identification of differentially expressed miRNAs (DeMiRs). The dataset included 1882 miRNAs, which were reduced to 220 miRNAs after normalization using the Limma package, R software. In the processing step using a heat map, the most important features were selected (Fig. 2B) and classified using machine learning algorithms. five algorithms (SVM, dts, rf, logistic regression, and knn) were then examined with four different metrics (accuracy, f1score, ROC_curve, and confusion matrix), and finally, according to the score obtained from these four metrics, the SVM algorithm was selected as the most accurate algorithm. (**DTS**, Accuracy: 88%, AUC = 47%; **RANDOM FOREST**, Accuracy: 93%, AUC = 39.5%; **SVM**, Accuracy: 93%, AUC = 88.5%; **KNN**, Accuracy: 93%, AUC = 41.7%; **LOGISTIC**, Accuracy: 93%, AUC = 88%). The confusion matrix can also be seen in Fig. 2C. As a result, a list of 29 miRNAs with five significant up and 24 significant down expressions in gastric cancer opted for further analysis (Table 2) Fig. 3.

ROC curve analysis for identification of diagnostic biomarkers. The findings of the ROC curve analysis demonstrated the diagnostic power of hsa-miR-29c (AUC of 0.7, with a sensitivity of 0.5 and specificity of 0.8, and cutoff of 0.88) which is improved when combined with hsa-miR-93 (combination had an AUC of 0.76, the sensitivity of 0.69 and specificity of 0.73 and cutoff of 0.86) (Fig. 2D).

Survival analysis of demiRs. Survival analysis of demiRs was performed using SPSS version 20, and the *p*-value was considered < 0.05 . the results demonstrated that 13 miRNAs (Hsa-miR-21, Hsa-miR-146b, Hsa-miR-185, Hsa-miR-1.1, Hsa-miR-1.2, Hsa-miR-143, Hsa-miR-4652, Hsa-miR-1911, Hsa-miR-29c, Hsa-miR-3170, Hsa-miR-139, Hsa-miR-5683, and Hsa-miR-133a.2) have the prognostic capability (Fig. 4).

Validation of candidate microRNAs in datasets. Among 29 candidate microRNAs resulted from machine learning algorithms, using online web servers mentioned in material and method section, the expression levels of hsa-miR-21, hsa-miR-133a, hsa-miR-146b, hsa-miR-29c, and hsa-miR-204 were highly validated in (EXP00118(GSE28700), EXP00131 (GSE23739), EXP00230 (GSE26595), EXP00268, EXP00270 (GSE54397) EXP00326 (GSE31568), EXP00337 (GSE59856), EXP00404, EXP00460 (GSE93415), EXP00524 (GSE106817), EXP00405, EXP00118(GSE28700), EXP00406, EXP00666, EXP00444 (GSE78775), EXP00476 (GSE99415), EXP00316 (GSE77380), and EXP00175(GSE33743)) supplementary file 1, heatmap analysis performed using miRPathDB online server (<https://mpd.bioinf.uni-sb.de/>), Fig. 5A).

miRNA-target prediction. The miRNA-target prediction was accomplished using several databases, that included: miRwalk, miRdb, and Targetscan. Using a Venn diagram online database, the list containing 407 common genes were identified (Fig. 5B).

Protein–protein interaction network analysis. Candidate genes predicted in the previous step were submitted to the STRING database to build a PPI network based on criteria mentioned in the material and method. To obtain hub genes with essential roles, the PPI network was then imported and visualized by Cytoscape software. 100 highly scored genes based on degree were selected using Cytoscape tools (Fig. 5C). Finally, the hub genes were imported into the string database to reanalyze the PPI network (Fig. 5D).

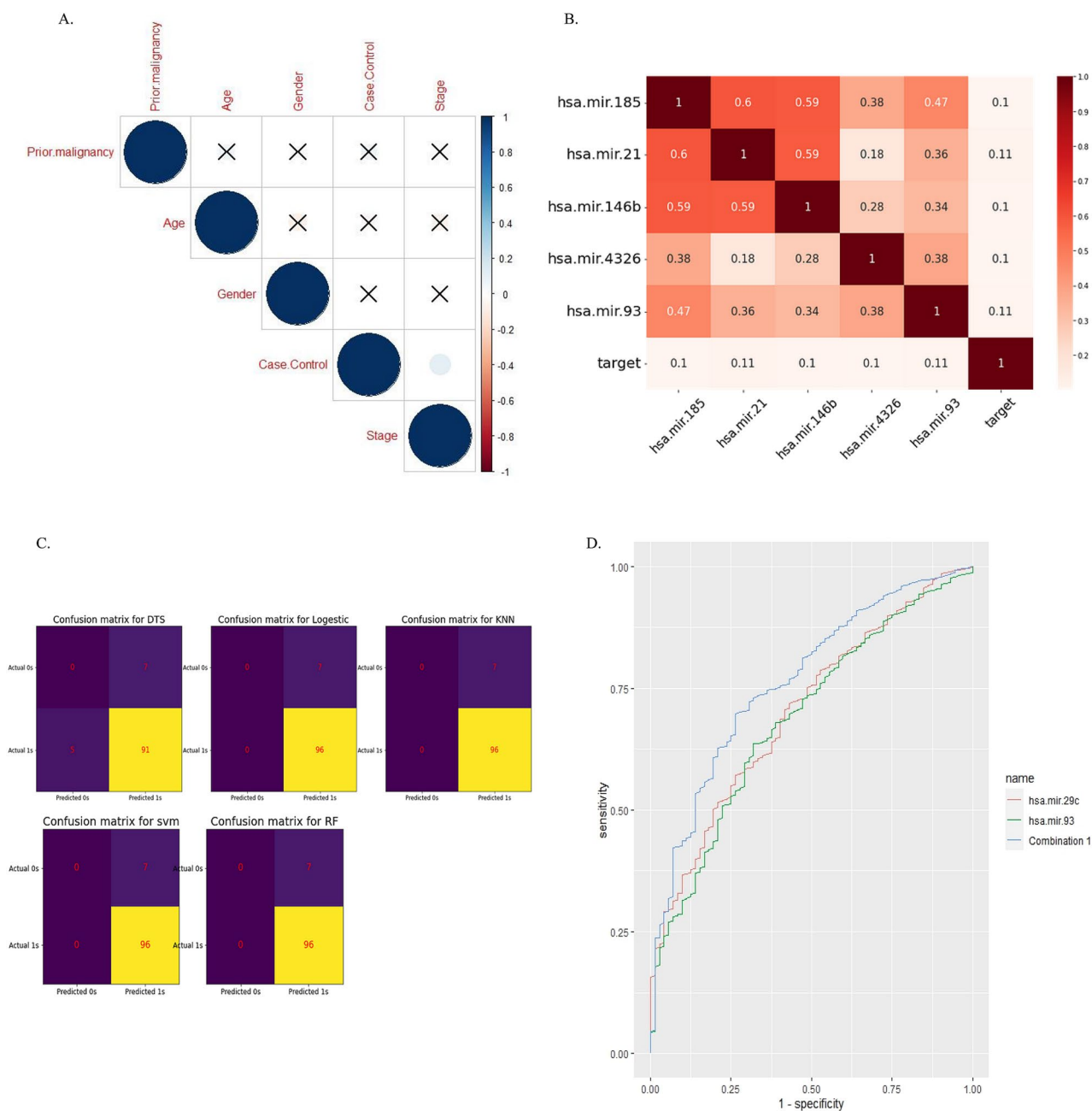


Figure 2. (A) correlation analysis using ggcorpot package, R software v 4.2; (B) Heat map analysis used to demonstrate important features. (C) The confusion matrix was used to compare the different machine-learning algorithms, figure B&C was plotted using Python v3.7. (D) ROC curve analysis revealed biomarker potency of miR-29c alone and in combination with miR-93 using R 4.2.2's combioROC package.

Functional analysis. To reveal the role of selected hub genes, enrichment analysis was undertaken using R software. The results demonstrated that hub genes transcription factor binding, enzyme binding, RNA polymerase II cis-regulatory region sequence-specific DNA binding, protein binding, double-stranded DNA binding, arrestin family protein binding, sequence-specific DNA binding, and chromatin binding, regarding molecular function, The majority of genes were enriched in miRNA mediated inhibition of translation, Positive regulation by host of viral transcription, Regulation of gene expression by genetic imprinting, Production of miRNAs involved in gene silencing by miRNA Gene silencing by miRNA, Wnt signaling pathway, calcium modulating pathway, Regulation of cellular senescence, Negative regulation of gene expression, and epigenetic Gene silencing concerning the biological process. Chromatin, Euchromatin, Nucleoplasm, Non-membrane-bounded organelle and Cytosol were top enriched cellular components. Identification of significant signaling pathways using the KEGG database shows that candidate hub genes were mainly involved in Glioma, Melanoma, Prostate cancer, Non-small cell lung cancer, Renal cell carcinoma, GnRH secretion, Aldosterone-regulated sodium reabsorption, and Pancreatic cancer (Fig. 6).

| | Fold change |
|----------------------|-------------|
| Upregulated miRNAs | |
| Hsa-miR-21 | 0.114061 |
| Hsa-miR-93 | 0.108749 |
| Hsa-miR-146b | 0.104449 |
| Hsa-miR-4326 | 0.1022 |
| Hsa-miR-185 | 0.100701 |
| Downregulated miRNAs | |
| Hsa-miR-6510 | -0.284227 |
| Hsa-miR-6507 | -0.227782 |
| Hsa-miR-6512 | -0.236185 |
| Hsa-miR-184 | -0.149492 |
| Hsa-miR-1.1 | -0.146 |
| Hsa-miR-1265 | -0.107646 |
| Hsa-miR-1266 | -0.107325 |
| Hsa-miR-143 | -0.103477 |
| Hsa-miR-4652 | -0.104071 |
| Hsa-miR-1911 | -0.106134 |
| Hsa-miR-29c | -0.142613 |
| Hsa-miR-3170 | -0.107759 |
| Hsa-miR-3622a | -0.140976 |
| Hsa-miR-378i | -0.127142 |
| Hsa-miR-4793 | -0.109031 |
| Hsa-miR-4461 | -0.114218 |
| Hsa-miR-204 | -0.130772 |
| Hsa-miR-139 | -0.137331 |
| Hsa-miR-551b | -0.137459 |
| Hsa-miR-205 | -0.125177 |
| Hsa-miR-5683 | -0.140301 |
| Hsa-miR-1.2 | -0.143739 |
| Hsa-miR-133a.1 | -0.107877 |
| Hsa-miR-133a.2 | -0.104760 |

Table 2. List of sign up and down-expressed miRNAs in gastric cancer.

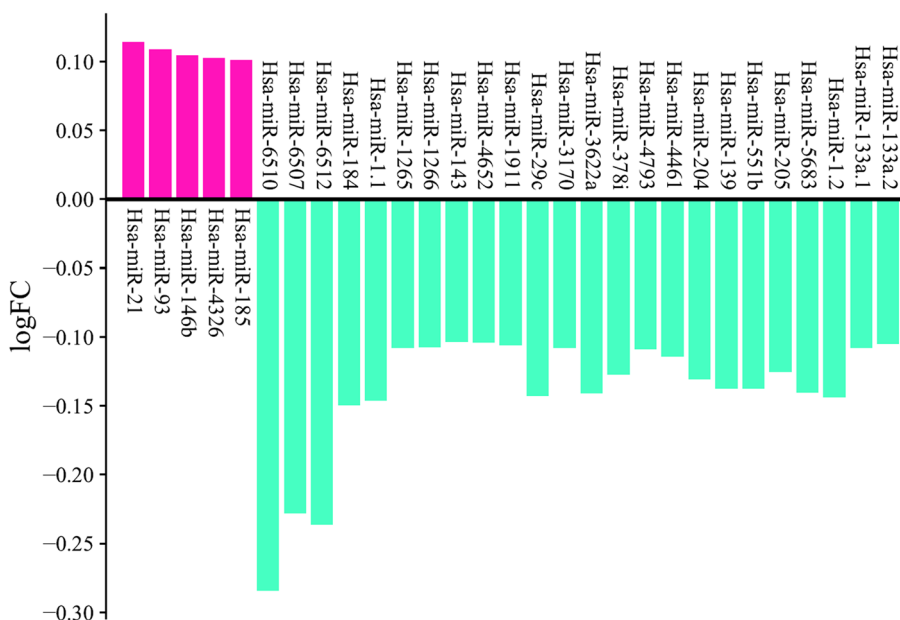


Figure 3. 29miRNAs obtained from feature selection-based machine learning based on logFC.

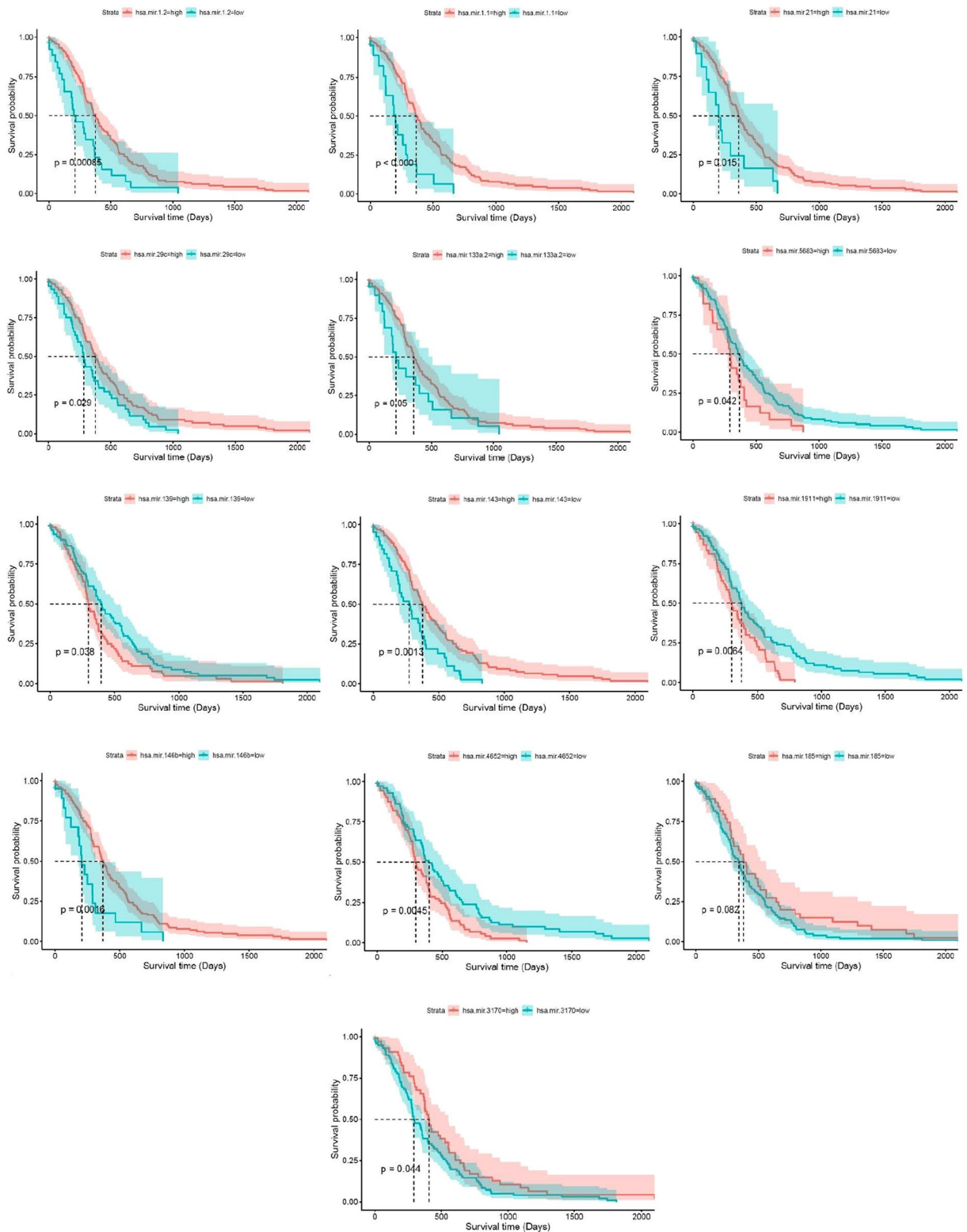


Figure 4. Kaplan Meier visualization of Identified prognostic biomarkers using Survival, survminer, and ggplot2 R packages from R software v4.2.2.

Discussion

Gastric cancer is a heterogeneous disease with a poor prognosis²⁸. Despite many advances in treatment including the use of adjuvant therapies such as surgery, chemotherapy, radiotherapy, and targeted therapy, the overall

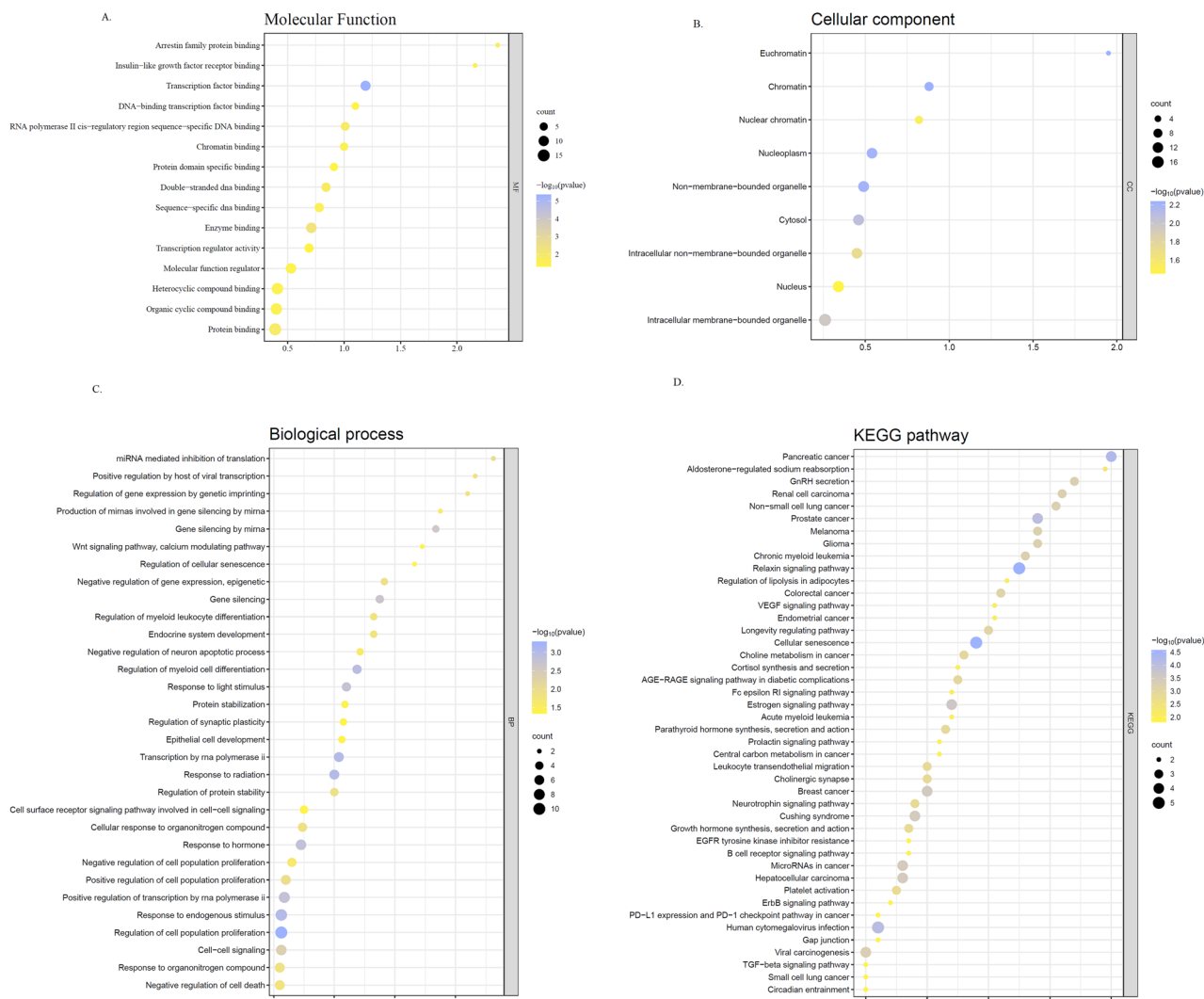


Figure 6. Functional enrichment analysis of target genes based on gene ontology regarding (A) molecular function, (B) cellular component (C) biological process. (D) KEGG pathway analysis. All the analysis was plotted using SRplot <http://www.bioinformatics.com.cn/srplot>, an online platform for data analysis and visualization.

5-year survival rate of the disease is still around 20–30% due to the diagnosis of patients in advanced stages of the disease²⁹. Therefore, it is important to identify biomarkers that can detect the disease in its early stages and also identify its likely prognosis. MicroRNAs (miRNAs) are a small member of the non-coding RNAs family that control various post-transcriptional processes and possess important role in carcinogenesis including angiogenesis, drug resistance, and cancer metastasis^{30,31}. In addition, the importance of miRNA-mRNA networks in activating or inhibiting many cancer-related molecular signaling pathways has recently been observed^{31,32}. Recently, with the development of bioinformatics as well as the aid of machine learning algorithms, different forms of next-generation sequencing are increasingly being used to detect biomarkers with a role in early detection, treatment, and prognosis of cancer^{33–35}. Machine learning approaches have gained popularity due to their predictive power of diagnostic and prognostic biomarkers. Studies also suggest the role of machine learning in feature selection such as miRNAs as an initial option when there is no prior knowledge^{36,37}. In the present study, TCGA data analysis of stomach cancer using machine learning algorithms identified 29 miRNA candidates. Even though machine learning offers different algorithms for selecting a gene as a biomarker, none of them alone and without clinical validation can guarantee the true power of a biomarker and the reproducibility of the results requires more investigations and validations³⁸. In order to overcome the limitations of machine learning, we validated the results obtained from it in different datasets such as GEO, TCGA, and GDAC, and among the 29 miRNAs found, hsa-miR-21, hsa-miR-133a, hsa-miR-146b, hsa-miR-29c, and hsa-miR-204 were among the most valid miRNAs in the investigated datasets.

Subsequent studies were designed to determine the role of these demiRs in gastric cancer through joint bioinformatics and prognosis analyses. In the first step, we explored the genes targeted by the five validated demiRs using online web server such as miRwalk, miRdb, and Targetscan. PPI network and cytohubba were then identified top 100 high-scored genes of each group of genes and functional and enrichment analyses were performed.

GO and KEGG analysis determined that DEGs were mainly involved in cancer-related pathways. In concert, it was demonstrated that these feature miRNAs are highly correlated with miRNA mediated inhibition of translation, Regulation of gene expression by genetic imprinting, Production of miRNAs involved in gene silencing by miRNA, Gene silencing by miRNA, Wnt signaling pathway, calcium modulating pathway, Regulation of cellular senescence, Negative regulation of gene expression based on functional enrichment analysis.

Survival analysis also reveals the prognostic capability of hsa-miR-21, hsa-miR-146b, hsa-miR-185, hsa-miR-1.1, hsa-miR-1.2, hsa-miR-143, hsa-miR-4652, hsa-miR-1911, hsa-miR-29c, hsa-miR-3170, hsa-miR-139, hsa-miR-5683, and hsa-miR-133a.2. On the other hand, ROC curve analysis demonstrated the diagnostic power of hsa-miR-29c in combination with hsa-miR-93. From these data, we can consider hsa-miR-21, hsa-miR-133a, hsa-miR-146b, and hsa-miR-29c a reliable biomarker panel. In a study by Larki et al., the expression level of miR-21 and miR-93 were investigated in GC patients with the aim of finding a panel of diagnostic signature, it was demonstrated that the over-expression of these miRNAs were highly correlated with early diagnosis and prognosis of this cancer³⁹. In another study conducted by CHAN et al. the diagnostic potency of miR-21 were detected, although they claimed that this miRNA can't serve as prognostic biomarker⁴⁰. Simonian considered miR-21 as both diagnostic and prognostic biomarker and suggested that it could be higher diagnostic potential compare to CA-19-9⁴¹. Also several study mentioned miR-133a as a tumor suppressor miRNA which inhibit the GC growth and exert its effect via different target genes such as Sp1, IGF1R, TCF4 and etc^{42–44}. Although the effect of hsa-miR-146b is less investigated in GC, in the study, the over-expression of this miRNA is demonstrated in GC patients, which exert its effect on NOVA1 and associated with poor prognosis⁴⁵. It is worth noting that based on the results of this study hsa-miR-29c have a dual biomarker potency and serves as both prognostic and diagnostic biomarker in GC. Although several studies mentioned the role of this miRNA as a diagnostic biomarker in GC, as far as our knowledge is concerned there isn't any study concerning its prognostic capability. On the other hand, the machine learning method used in this study was demonstrated three significant down-regulated miRNAs hsa-miR-6510, hsa-miR-6507, hsa-miR-6512 in gastric cancer which recently gained attention in gastric cancer. Li et al. showed that miR-6512 downregulated in gastric cancer patients and seems to have a correlation with skin manifestations and fibrosis⁴⁶. Moreover, a study by Ding demonstrated that miR-6507 was among top five down expressed miRNAs in GC and may serve as a good predictive biomarker⁴⁷. Another study introduced miR-6510 as a potential prognostic biomarker in GC patients⁴⁸. Although their role is not fully identified and need further investigation, these miRNAs could play an important role in GC.

Conclusion

In conclusion, 29 miRNAs were identified using machine learning algorithms, subsequent analyzes showed a panel of four miRNAs including hsa-miR-21, hsa-miR-133a, hsa-miR-146b, and hsa-miR-29c with high diagnostic and prognostic power, which was validated by several datasets. This study emphasizes the importance of machine learning as an alternative option for predicting biomarkers in gastric cancer. However, for the purpose of further validation, it is suggested that candidate miRNAs will be analyzed in several cohorts in a laboratory manner.

Data availability

The datasets used and/or analyzed during the present study are available from the corresponding author upon reasonable request.

Received: 2 December 2022; Accepted: 26 March 2023

Published online: 15 April 2023

References

- Zhang, Z. et al. A five-miRNA signature predicts survival in gastric cancer using bioinformatics analysis. *Gene* **699**, 125–134. <https://doi.org/10.1016/j.gene.2019.02.058> (2019).
- Guan, Y.-J., Ma, J.-Y. & Song, W. Identification of circRNA–miRNA–mRNA regulatory network in gastric cancer by analysis of microarray data. *Cancer Cell Int.* **19**(1), 183. <https://doi.org/10.1186/s12935-019-0905-z> (2019).
- Wu, F. et al. Network analysis based on TCGA reveals hub genes in colon cancer. *Contemp. Oncol.* **21**(2), 136. <https://doi.org/10.5114/wo.2017.68622> (2017).
- Shi, X.-H. et al. A five-microRNA signature for survival prognosis in pancreatic adenocarcinoma based on TCGA data. *Sci. Rep.* **8**(1), 1–10. <https://doi.org/10.1038/s41598-018-22493-5> (2018).
- Huynh-Thu, V. A. et al. Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. *Bioinformatics* **28**(13), 1766–1774. <https://doi.org/10.1093/bioinformatics/bts238> (2012).
- Gilani, N. et al. Identifying potential miRNA biomarkers for gastric cancer diagnosis using machine learning variable selection approach. *Front. Genet.* **12**, 2538. <https://doi.org/10.3389/fgene.2021.779455> (2021).
- Rehman, O. et al. Validation of miRNAs as breast cancer biomarkers with a machine learning approach. *Cancers* **11**(3), 431. <https://doi.org/10.3390/cancers11030431> (2019).
- Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The cancer genome atlas (TCGA): An immeasurable source of knowledge. *Contemp. Oncol. (Pozn)* **19**(1a), A68–77. <https://doi.org/10.5114/wo.2014.47136> (2015).
- Greener, J. G. et al. A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* **23**(1), 40–55. <https://doi.org/10.1038/s41580-021-00407-0> (2022).
- Rajkomar, A., Dean, J. & Kohane, I. Machine learning in medicine. *N. Engl. J. Med.* **380**(14), 1347–1358. <https://doi.org/10.1056/NEJMra1814259> (2019).
- Xu, A., Kouznetsova, V. L. & Tsigelny, I. F. Alzheimer's disease diagnostics using miRNA biomarkers and machine learning. *J. Alzheimer's Dis.* <https://doi.org/10.3233/JAD-215502> (2022).
- Lee, J. Y. et al. Radiomic machine learning for predicting prognostic biomarkers and molecular subtypes of breast cancer using tumor heterogeneity and angiogenesis properties on MRI. *Eur. Radiol.* **32**(1), 650–660. <https://doi.org/10.1007/s00330-021-08146-8> (2022).

13. Maurya, N. S. *et al.* Transcriptome profiling by combined machine learning and statistical R analysis identifies TMEM236 as a potential novel diagnostic biomarker for colorectal cancer. *Sci. Rep.* **11**(1), 1–11. <https://doi.org/10.1038/s41598-021-92692-0> (2021).
14. Mahesh, B. Machine learning algorithms—a review. *Int. J. Sci. Res. (IJSR)* **9**, 381–386. <https://doi.org/10.21275/ART20203995> (2020).
15. Bijalwan, V. *et al.* KNN based machine learning approach for text and document mining. *Int. J. Database Theory Appl.* **7**(1), 61–70. <https://doi.org/10.14257/ijdata.2014.7.1.0> (2014).
16. Rymarczyk, T. *et al.* Logistic regression for machine learning in process tomography. *Sensors* **19**(15), 3400. <https://doi.org/10.3390/s19153400> (2019).
17. Macaulay, M. O. & Shafiee, M. Machine learning techniques for robotic and autonomous inspection of mechanical systems and civil infrastructure. *Auton. Intell. Syst.* **2**(1), 1–25. <https://doi.org/10.1007/s43684-022-00025-3> (2022).
18. Saseendran, A. T. *et al.* Impact of noise in dataset on machine learning algorithms. In: *Machine Learning Research*, 0–8. (2019). <https://doi.org/10.13140/RG.2.2.25669.91369>.
19. Bertoli, G. D. C. *et al.* An end-to-end framework for machine learning-based network intrusion detection system. *IEEE Access* **9**, 106790–106805. <https://doi.org/10.1109/ACCESS.2021.3101188> (2021).
20. Markoulidakis, I. *et al.* Multiclass confusion matrix reduction method and its application on net promoter score classification problem. *Technologies* **9**(4), 81. <https://doi.org/10.3390/technologies9040081> (2021).
21. Sticht, C. *et al.* miRWalk: an online resource for prediction of microRNA binding sites. *PLoS ONE* **13**(10), e0206239. <https://doi.org/10.1371/journal.pone.0206239> (2018).
22. Chen, Y. & Wang, X. miRDB: An online database for prediction of functional microRNA targets. *Nucleic Acids Res.* **48**(D1), D127–D131. <https://doi.org/10.1093/nar/gkz757> (2020).
23. McGeary, S. E. *et al.* The biochemical basis of microRNA targeting efficacy. *Science* **366**(6472), eaav1741. <https://doi.org/10.1126/science.aav1741> (2019).
24. Szklarczyk, D. *et al.* The STRING database in 2021: Customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**(D1), D605–D612. <https://doi.org/10.1093/nar/gkaa1074> (2020).
25. Consortium TGO. The Gene Ontology resource: Enriching a gold mine. *Nucleic Acids Res.* **49**(D1), D325–D334. <https://doi.org/10.1093/nar/gkaa1113> (2020).
26. Kanehisa, M. *et al.* KEGG: Integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**(D1), D545–d551. <https://doi.org/10.1093/nar/gkaa970> (2021).
27. Shannon, P. *et al.* Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**(11), 2498–2504. <https://doi.org/10.1101/gr.1239303> (2003).
28. Joshi, S. S. & Badgwell, B. D. Current treatment and recent progress in gastric cancer. *CA Cancer J. Clin.* **71**(3), 264–279 (2021).
29. Thrift, A. P. & El-Serag, H. B. Burden of gastric cancer. *Clin. Gastroenterol. Hepatol.* **18**(3), 534–542. <https://doi.org/10.1016/j.cgh.2019.07.045> (2020).
30. Naorem, L. D., Muthaiyan, M. & Venkatesan, A. Identification of dysregulated miRNAs in triple negative breast cancer: A meta-analysis approach. *J. Cell. Physiol.* **234**(7), 11768–11779. <https://doi.org/10.1002/jcp.27839> (2019).
31. Samami, E. *et al.* The potential diagnostic and prognostic value of circulating MicroRNAs in the assessment of patients with prostate cancer: Rational and progress. *Front. Oncol.* **11**, 5977. <https://doi.org/10.3389/fonc.2021.716831> (2021).
32. Ding, B. *et al.* Whole-transcriptome analysis reveals a potential hsa_circ_0001955/hsa_circ_0000977-mediated miRNA-mRNA regulatory sub-network in colorectal cancer. *Aging (Albany NY)* **12**(6), 5259. <https://doi.org/10.18632/aging.102945> (2020).
33. Iqbal, M. J. *et al.* Clinical applications of artificial intelligence and machine learning in cancer diagnosis: Looking into the future. *Cancer Cell Int.* **21**(1), 270. <https://doi.org/10.1186/s12935-021-01981-1> (2021).
34. Nazari, E. *et al.* Machine learning approaches for classification of colorectal cancer with and without feature selection method on microarray data. *Gene Rep.* **25**, 101419. <https://doi.org/10.1016/j.genrep.2021.101419> (2021).
35. Nazari, E. *et al.* Identification of potential biomarkers in stomach adenocarcinoma using machine learning approaches. *Curr. Bioinform.* <https://doi.org/10.2174/1574893618666230227103427> (2023).
36. Ma, J. *et al.* Bioinformatic analysis reveals an exosomal miRNA-mRNA network in colorectal cancer. *BMC Med. Genom.* **14**(1), 1–18 (2021).
37. Nazari Elham*, P. G., Asadnia Alireza, Dashtiahangar Mohammad, Mohit Reza, Khazaei Majid*, Maftooh Mina, Nassiri Mohamadreza, Hassanian Seyed Mahdi, Ghayour-Mobarhan Majid, Ferns A Gordon, Shahidsales Soodabeh and Avan Amir*, Identification of potential biomarkers in stomach adenocarcinoma using machine learning approaches. *Curr. Bioinform.* **18**, <https://doi.org/10.2174/1574893618666230227103427> (2023).
38. Kim, A. A., Rachid Zaim, S. & Subbian, V. Assessing reproducibility and veracity across machine learning techniques in biomedicine: A case study using TCGA data. *Int. J. Med. Inform.* **141**, 104148. <https://doi.org/10.1016/j.ijmedinf.2020.104148> (2020).
39. Larki, P. *et al.* Up-Regulation of miR-21, miR-25, miR-93, and miR-106b in gastric cancer. *Iran. Biomed. J.* **22**(6), 367–373. <https://doi.org/10.29252/ibj.22.6.367> (2018).
40. Chan, S.-H. *et al.* miR-21 microRNA expression in human gastric carcinomas and its clinical association. *Anticancer Res.* **28**(2A), 907–911 (2008).
41. Simonian, M., Mosallayi, M. & Mirzaei, H. Circulating miR-21 as novel biomarker in gastric cancer: Diagnostic and prognostic biomarker. *J. Cancer Res. Ther.* **14**(2), 475 (2018).
42. Qiu, T. *et al.* MiR-145, miR-133a and miR-133b inhibit proliferation, migration, invasion and cell cycle progression via targeting transcription factor Sp1 in gastric cancer. *FEBS Lett.* **588**(7), 1168–1177 (2014).
43. He, M.-Q. *et al.* miR-133a-5p suppresses gastric cancer through TCF4 down-regulation. *J. Gastrointest. Oncol.* **12**(3), 1007 (2021).
44. Gong, Y. *et al.* Tumor suppressor role of miR-133a in gastric cancer by repressing IGF1R. *World J. Gastroenterol. WJG* **21**(10), 2949 (2015).
45. Yoon, S. O. *et al.* NOVA1 inhibition by miR-146b-5p in the remnant tissue microenvironment defines occult residual disease after gastric cancer removal. *Oncotarget* **7**(3), 2475–2495. <https://doi.org/10.18632/oncotarget.6542> (2016).
46. Li, X. *et al.* Novel CircRNAs in Hub ceRNA axis regulate gastric cancer prognosis and microenvironment. *Front. Med.* <https://doi.org/10.3389/fmed.2021.771206> (2021).
47. Ding, W. *et al.* Comprehensive analysis of competitive endogenous RNAs network: Identification and validation of prediction model composed of mRNA signature and miRNA signature in gastric cancer. *Oncol. Lett.* **23**(5), 150. <https://doi.org/10.3892/ol.2022.13270> (2022).
48. Woo, Y. *et al.* Tumor epigenetic signature and survival in resected gastric cancer patients. *J. Am. Coll. Surg.* **232**(4), 483–491.e1. <https://doi.org/10.1016/j.jamcollsurg.2020.12.023> (2021).

Acknowledgements

The authors acknowledge the financial support from Mashhad University of Medical Sciences.

Author contributions

H.A., E.N.: the conception and design of the study, or acquisition of data, or analysis and interpretation of data; R.M., A.A.: acquisition of data and analysis; M.K., S.M.H., M.M., M.G.-M., G.A.F., S.S., M.N., and A.A. have provided critical revision of the final manuscript. E.N. and A.A. have contributed to the study design and approved the final version of the manuscript. All the authors read and approved the final version of the manuscript.

Funding

This study was supported by a grant awarded to Amir Avan by the Mashhad University of Medical Sciences.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-32332-x>.

Correspondence and requests for materials should be addressed to G.A.F. or A.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023