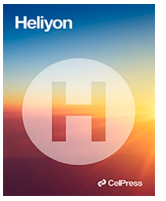
Heliyon 11 (2025) e41443

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Development and validation of a biomarker-based prediction model for metastasis in patients with colorectal cancer: Application of machine learning algorithms

Erfan Ayubi [a], Sajjad Farashi [b], Leili Tapak [c,d], Saeid Afshar [a,e,*]

[a] *Cancer Research Center, Institute of Cancer, Avicenna Health Research Institute, Hamadan University of Medical Sciences, Hamadan, Iran*
[b] *Neurophysiology Research Center, Institute of Neuroscience and Mental Health, Avicenna Health Research Institute, Hamadan University of Medical Sciences, Hamadan, Iran*
[c] *Modeling of Noncommunicable Diseases Research Center, Institute of Health Sciences andTechnologies, Avicenna Health Research Institute, Hamadan University of Medical Sciences, Hamadan, Iran*
[d] *Department of Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran*
[e] *Department of Medical Biotechnology, School of Advanced Medical Sciences and Technologies, Hamadan University of Medical Sciences, Hamadan, Iran*

A R T I C L E I N F O

Keywords:
Colorectal cancer
Metastasis
Machine learning
Biomarker

A B S T R A C T

Objective: The purpose of the current study was to develop and validate a biomarker-based prediction model for metastasis in patients with colorectal cancer (CRC).
Methods: Two datasets, GSE68468 and GSE41568, were retrieved from the Gene Expression Omnibus (GEO) database. In the GSE68468 dataset, key biomarkers were identified through a screening process involving differential expression analysis, redundancy analysis, and recursive feature elimination technique. Subsequently, the prediction model was developed and internally validated using five machine learning (ML) algorithms including lasso and elastic-net regularized generalized linear model (glmnet), k-nearest neighbors (kNN), support vector machine (SVM) with Radial Basis Function Kernel, random forest (RF), and eXtreme Gradient Boosting (XGBoost). The predictive performance of the algorithm with the highest accuracy was then externally validated on the GSE41568 dataset.
Results: Among 22,283 registered genes in the GSE68468 dataset, the screening process identified 16 key genes including *MMP3, CCDC102B, CDH2, SCGB1A1, KRT7, CYP1B1, LAMC3, ALB, DIXDC1, VWF, MMP1, CYP4B1, NKX3-2, TMEM158, GADD45B, SERPINA1* and these genes were used to build the prediction model. On the internal validation dataset, the prediction performance of five ML algorithms was as follows; RF (accuracy = 0.97 and kappa = 0.91), XGBoost (0.93, 0.81), kNN (0.93, 0.81), glmnet (0.93, 0.82) and SVM (0.92, 0.80). Top five biomarkers were *MMP3, CCDC102B, CDH2, VWF* and *MMP1*. The RF model exhibited an accuracy of 0.97, a kappa value of 0.92, and an area under the curve (AUC) of 0.99 in the external validation dataset.
Conclusion: The results of this study have identified biomarkers through ML algorithms which help to identify patients with CRC prone to metastasis.

* Corresponding author. Cancer Research Center, Institute of Cancer, Avicenna Health Research Institute, Hamadan University of Medical Sciences, Hamadan, Iran.
E-mail address: safshar.h@gmail.com (S. Afshar).

https://doi.org/10.1016/j.heliyon.2024.e41443
Received 2 March 2024; Received in revised form 21 December 2024; Accepted 22 December 2024
Available online 24 December 2024
2405-8440/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

An approximate incidence of 1.9 million and around 0.9 million fatalities related to colorectal cancer (CRC) were observed worldwide in 2020 [1]. It is also estimated that CRC incidence and mortality will increase by more than 1.5 times by 2040 [1]. The 5-year survival rate and prognosis for individuals with metastatic CRC is exceedingly low [2]. The poor prognosis is mostly due to the development of metastases in vital organs (e.g., liver and lung) [3], intestinal obstruction [4], and the presence of metastases at the time of diagnosis [5].

In this context, identifying susceptible patients likely to progress to metastasis using risk prediction models is crucially important. It can lead to the early detection of metastasis-prone CRC tumors, guide the treatment modalities, and improve survival rate [6]. Previous risk prediction models show that CRC metastasis may be a function of histopathological and biomarker risk factors [7]. Although an umbrella review of all systematic reviews showed the evidence credibility of some histopathological risk factors for CRC metastasis was graded as highly suggestive, the credibility level of evidence for biomarkers is still weak [7].

This level of evidence conveys that multiple and larger studies are needed to explore the statistically significant effect of biomarker-based predictive models for CRC metastasis. Advancements in bioinformatics methods enable researchers to study the role of molecular markers in predicting and diagnosing CRC cancer outcomes [8,9]. Although numerous studies have identified single gene biomarkers for CRC metastases with sufficient prediction and diagnosis performance metrics [10–13], there is an argument that efforts should be directed towards working with multi-omics and high-dimensional dataset to identify multiple gene markers [14]. Some studies have presented polygenic risk prediction models for CRC outcomes using the high-dimensional datasets available in the Cancer Genome Atlas (TCGA) and the Gene Expression Omnibus (GEO) databases [15–17]. Although the aforementioned studies provide valuable information on multi-gene models for predicting the outcomes in patients with CRC, however, some of them have pointed to the limited number of metabolic genes to evaluate the prognostic model and also the small size of validation datasets as limitations [16, 17]. Another important note that should considered is that although using narrow-down approaches to a small subset of biomarkers by using stringent thresholds will always yield a panel of biomarkers, having multiple genes prediction models with a large number of biomarkers may still lead to complexity in the model. Here, there is a need to make a trade-off between simplicity and complexity and to prioritize the biomarkers in the model in terms of importance.

Identifying important differentially expressed genes (DEGs) based on multi-omics big data requires powerful tools that can accurately analyze the importance of biomarkers, predict the outcomes, and also classify groups. Machine learning (ML) algorithms can handle multi-omics big data and discover robust and reproducible predictive biomarkers [18].

In previous studies [19–21], the performance of linear, non-linear, and boosting ML algorithms considering demographic and clinical data as predictors for diagnosis and prediction of outcomes in CRC patients have been investigated, yielding a high accuracy rate in clinical practice. It has been argued that the gradient boosting machine (GBM) algorithms (e.g., eXtreme Gradient Boosting, XGBoost) outperform other classical algorithms such as random forest (RF), support vector machine (SVM) and k-nearest neighbors (kNN) in cancer prediction and prognosis [22,23].

Although the evidence about developing and validating the prediction models utilizing demographic and clinical data as predictors is sufficient and credible in CRC research, further studies are needed to assess the ability of gene signatures for metastasis prediction from primary cancer tissues [24]. Furthermore, the accuracy and interpretability of ML-based models for metastasis prediction are highly sensitive to the selection of the most important features. In the context of cancer diagnosis and prediction, Fold-change (FC) cut-off, student's t-test, or ANOVA are common methods for selecting DEGs from microarray data [25]. These commonly used feature selection methods may be associated with inherent problems and limit confidence in the results of prediction models [25,26]. A popular automatic method for feature selection is traditional Recursive Feature Elimination (RFE). It is a backward selection approach and is based on the learned model and classification accuracy. The new RFE technique can be coupled with other classification models such as SVM, RF, and GBM [27].

It is hypothesized that ML algorithms based on a hierarchical feature selection process which consists of FC cut-off, redundancy analysis, and RFE algorithm could provide accurate predictions. Considering the aforementioned issues, the current study aimed to develop a biomarker-based predictive model for metastasis prediction in patients with CRC utilizing classic ML algorithms and a variant of GBM based on a hierarchical feature selection process. To assess the generalizability of the model, an internal-external validation analysis was also performed.

2. Methods

2.1. Dataset and patients

We utilized two gene expression datasets, i.e. GSE68468 (with the platform GPL96) and GSE41568 (with the platform GPL570) available from the international and public-use functional genomics data repository, GEO database [28]. The two datasets are accessible through NCBI GEO (https://www.ncbi.nlm.nih.gov/geo). The main criteria for selecting datasets from the GEO database were: The dataset should include both groups of patients with metastatic and primary CRC tumors, the number of samples in both groups (at least 50 subjects), and also overlapping of the two datasets in terms of genes.

The GSE68468 involves 390 samples. After excluding non-CRC samples (n = 128), the GSE68468 dataset comprised 262 samples with CRC, including 67 metastatic samples and 195 non-metastatic samples according to histology. Similarly, the GSE41568 dataset included 133 samples with CRC, with 94 metastatic cases and 39 non-metastatic cases.

A biomarker-based predictive model was trained and tested using the GSE68468 dataset. The external validity of the developed

model was evaluated in the GSE41568 dataset. Details of the number of subjects studied in two datasets, GSE68468 and GSE41568, including the number of patients with CRC, those without CRC, and also a number of malignant and non-malignant cases, are shown in the **Supplementary file (**Fig. S1).

2.2. Differential expression analysis, redundancy and importance analyses and feature selection

The GEO2R online tool was applied to identify potential important DEGs between non-metastatic and metastatic samples. The threshold was $|\log(FC)| > 1$ (base 2 logarithm of FC) and Benjamini-Hochberg adjusted P-value less than 0.05. The conservative log FC cutoff of >1 was used to preserve all potentially significant genes for further analysis. The performance and generalizability of ML algorithms could be improved if highly correlated features are not present in the model [29]. The correlation matrix using the findCorrelation function in the Classification And REgression Training (caret) package in R was created to identify duplicate or highly correlated genes. If two genes are highly correlated, the function takes into account the mean absolute correlation of each gene and removes the one with the largest mean absolute correlation. The genes with an absolute pair-wise correlation of 0.70 or higher were removed. With doing this, it is expected that among two highly correlated genes that shared substantial variance (e.g., 50 %), one representative gene will be for further analysis and also the size of the dataset has effectively reduced for downstream analyses [29]. After discarding redundant genes, the RFE algorithm was applied for gene selection. The RFE works in several steps. At first, the genes are ranked based on their importance, and less important genes are removed, and a model is built with the remaining genes. These steps continue until the optimal number of genes is reached [30,31].

The RFE algorithm can be used with any of the ML algorithms. There are several algorithms for RFE through caret package of R software. The random forest (RF) algorithm called rfFuncs was accordingly used. Furthermore, a 10-fold cross-validation with 3 repeats was used to evaluate the generalizability of the RFE algorithm for feature selection. Due to the randomness of the results, algorithm RFE was repeated 10 times. In each repetition, genes with higher ranks in terms of influence on the accuracy were identified. After the 10 repetitions, those that appeared most frequently in the collection of influential genes were considered predictors in ML algorithms.

2.3. Data Pre-Processing, machine learning algorithms and hyperparameter optimization

We transformed the data to enhance the ability of ML algorithms to improve the numerical stability and yield the best results. We used the basic data transforms of center and scale using the caret package. In this transformation, each gene is distributed with a mean of 0 and a standard deviation of 1. Five different ML algorithms were used, including lasso and elastic-net regularized generalized linear model (glmnet), kNN, SVM with Radial Basis Function Kernel, RF, and XGBoost, using key identified important genes from the RFE algorithm.

The important parameters with a substantial effect on ML algorithms were tuned using a grid search, which involves exploring a grid of parameters to find the optimal values for each model parameter. The optimization was performed for RF by tuning the number of features randomly sampled at each split (mtry), glmnet with the elasticnet mixing parameter and lambda sequence, kNN with the number of neighbors, SVM with regularization parameter and the distance a single training example reaches, and finally XGboost with number of trees, maximum tree depth, learning rate, regularization parameter, percentage of columns to build each tree, the minimum weight required to create a new node in the tree, and subsampling proportion of training instances.

2.4. Training and testing

The GSE68468 dataset was split into two parts, with 70 % used for training the models, and the remaining 30 % for internal validation. The 10-fold cross-validation (CV) was used to estimate the accuracy of the ML algorithm on the first part of the dataset. In other words, the training dataset is split into 10 parts, train in 9, and test on 1. In this method, during 10 iterations, each time one of the 10 parts is considered as the test set, and the remaining 9 parts are considered as the training set. Finally, the mean of the results in the iterations is considered an overall accuracy estimate. The process of splitting the dataset into 10-fold CV is repeated 3 times. The mean of 3 repeats was considered as the final model accuracy.

2.5. Model Evaluation metrics and importance score normalization

The power of ML algorithms to predict CRC metastasis in the training dataset was evaluated

using accuracy, kappa, and the area under the curve (AUC) metrics. Accuracy represents the percentage of correctly classified patients out of all patients, while kappa is similar to accuracy but accounts for random chance. AUC provides an overall measure of the performance of the model for classification and discrimination.

In the internal and external validation datasets, ML performance was evaluated using the following measures: accuracy, kappa, specificity, precision, recall, and F1 score. The degree of usefulness of each gene in the model prediction was determined using estimation of importance score. The importance score of biomarkers from best ML algorithm have been rescaled using min-max normalization. This normalization is a linear transformation on the dataset so that the relationship among data values is also preserved. With doing normalization, the minimum importance score have been subtracted from each importance score and then divide the result by the difference between the maximum importance score and the minimum importance score. Thus, the rescaled scores fall between 0 and 1.

Due to the expected randomness of ML results, all ML algorithms were repeated 10 times to reduce the variability introduced by random splitting, and then a summary of the distribution of performance metrics (mean and standard deviation) was reported.

The ML algorithms were implemented using the caret, glmnet, and xgboost packages in R version 4.1.3. The Mann-Whitney *U* Test was used to compare differences in the median of genes between metastatic and non-metastatic CRC patients. The P-value<0.05 was considered as statistical significance. R-codes used for analysis is provided in the supplementary file (Source code).

3. Results

3.1. Identification of important metastasis-associated genes

As shown in Fig. 1, the GSE68468 dataset initially involved 22,283 genes, of which 21,948 were removed using the statistical criteria i.e. |log(FC)| > 1, and adjusted P-value <0.05. From the remaining 335 genes, 14 were without an identifiable symbol and were discarded. Applying a cutoff of ≥|0.70| for the correlation coefficient, 197 genes were suspected of being redundant and were subsequently discarded. Out of the remaining 124 genes, the RFE algorithm demonstrated high accuracy and kappa up to the number of 16 genes. Beyond this gene number, there was no noticeable change in accuracy and kappa. The genes that appeared most frequently in the collection of the top 16 genes in the 10 repetitions of algorithm RFE were as follows; *MMP3, CCDC102B, CDH2, SCGB1A1, KRT7, CYP1B1, LAMC3, ALB, DIXDC1, VWF, MMP1, CYP4B1, NKX3-2, TMEM158, GADD45B, SERPINA1*. These 16 genes were considered predictors for ML algorithms.

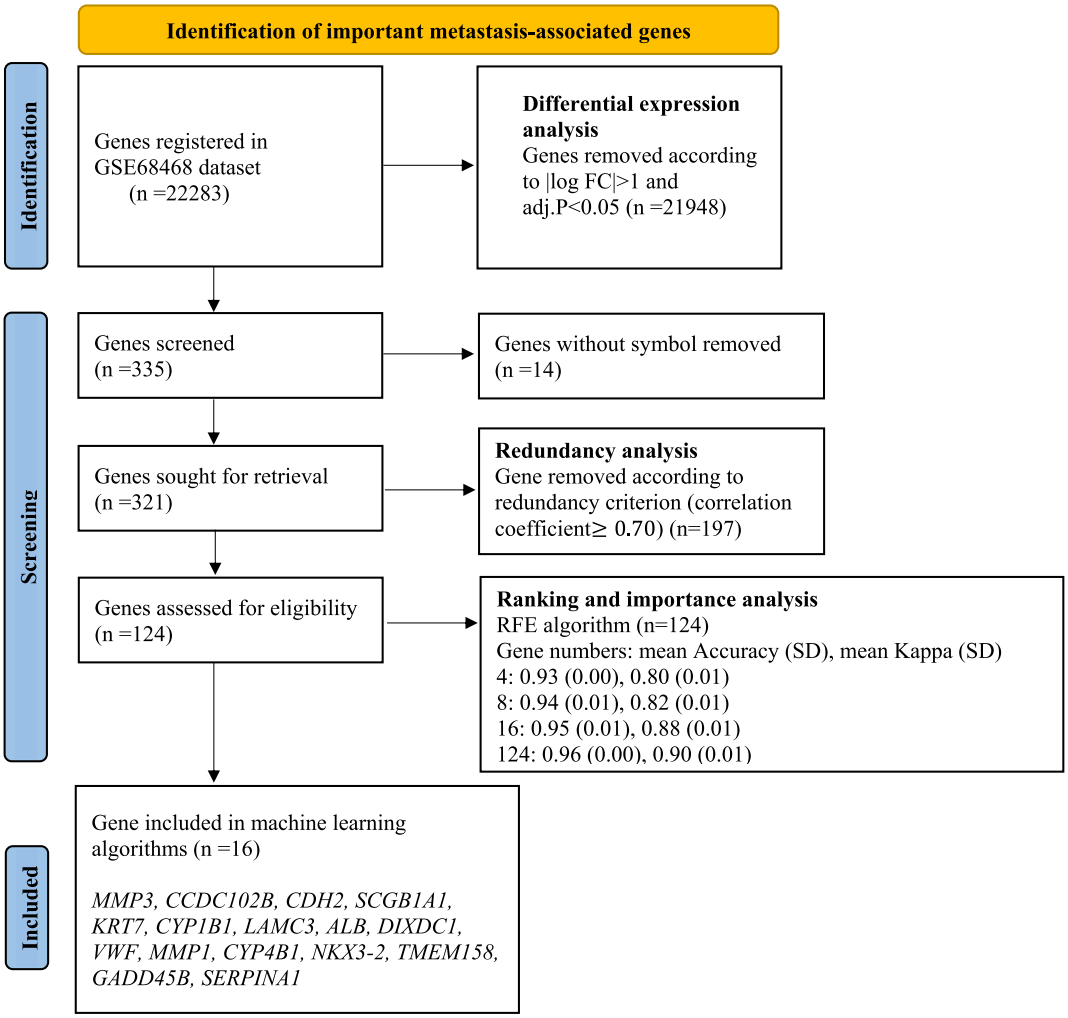


Fig. 1. Flow diagram of the biomarker screening and identification workflow.

3.2. Development and validation of the model using the internal dataset

The model performance of the five ML algorithms in the internal training dataset is presented in Table 1. From the mean accuracy measure viewpoint, ML algorithms can be arranged as, RF, kNN, SVM, glmnet, and XGBoost respectively. The negligible standard deviation of reported values for different runs of algorithms indicated the robustness of these results. The RF model (mean accuracy = 0.97 and mean kappa = 0.91) outperformed other models. Additionally, RF achieved a 99 % AUC. Characteristics of 16 important genes, as well as their expression patterns according to metastatic and non-metastatic CRC, are presented in Table 2. *MMP1* and *SCGB1A1* had the highest and lowest |log(FC)| values, respectively. The median expression pattern of the identified genes was statistically different between non-metastatic and metastatic CRC (p-value <0.05). After normalizing importance scores from the RF model as the algorithm with the best performance to the 0–1 range, *MMP3* (importance score = 1), *CCDC102B* (0.80), *CDH2* (0.49), *VWF* (0.31), and *MMP1* (0.27) were the top five biomarkers for predicting CRC metastasis. According to Table 2, for the first 5 important biomarkers (based on importance score), the expression of *MMP3*, *VWF* and *MMP1* in the non-metastatic CRC group was significantly higher as compared with metastatic samples. While for *CCDC102B* and *CDH2* biomarkers, the expression was significantly higher in metastatic CRC samples.

The 2D scatter plot of the top five biomarkers to differentiate non-metastatic and metastatic CRC in all subjects of the GSE68468 dataset is shown in Fig. 2. Since the distribution of the biomarkers deviates from the normal distribution, each biomarker data value is replaced with logarithm base 10. The compact and well-separated clouds in the feature space with a minimal overlap indicated the discriminant capability of these genes. As is intuitively clear, the pairs of log*MMP3* and the other four biomarkers including log*CCDC102B*, log*CDH2*, log*VWF*, and log*MMP1* have better discrimination to differentiate non-metastatic and metastatic CRC compared with other pairs. In the internal testing dataset, the RF model showed the best performance among other algorithms (accuracy = 0.97, kappa = 0.91, specificity = 0.98, precision = 0.95, recall = 0.92, and F1 score = 0.93). (Table 3).

3.3. Validation using external independent dataset

The performance of the proposed prediction model based on the RF algorithm in the GSE41568 dataset is depicted as a fourfold plot (Fig. 3). Ninety-seven percent of patients were correctly classified (mean accuracy = 0.97). Other performance measures were as follows: kappa = 0.92, specificity = 0.97, recall = 0.97, precision = 0.99, and F1 score = 0.98. Moreover, the average AUCs over 10 repetitions was 0.99.

4. Discussion

The purpose of our study was to develop and validate an ML-based biomarker predictive model for metastasis in patients with CRC using GEO data. The training and testing of ML algorithms revealed that the RF model had the best performance compared to other algorithms, achieving high accuracy rates of 0.98 and 0.96 in the internal and external validation datasets, respectively. Important biomarkers for the prediction of metastases were *MMP3*, *CCDC102B*, *VWF*, *MMP1*, and *CDH2*, respectively.

MMP3, as a member of the zinc-dependent endopeptidases family, plays an essential role in tumor progression and metastasis [32]. The protein encoded by this gene regulates cell invasion and angiogenesis by proteolysis of different proteins such as other MMPs, different types of collagens, fibronectin, elastin, proteoglycans, and adhesion molecules [33]. Liang et al. [34], in their study, demonstrated that *MMP3* is upregulated in several tumors and associated with tumor growth and metastasis. Despite the potential role

Table 1
Model performance of five ML algorithms in the internal training dataset.

Accuracy	Min	1st Q	Median	Mean	3rd Q	Max
glmnet	0.81 (0.05)	0.93 (0.02)	0.96 (0.02)	0.95 (0.01)	0.98 (0.03)	1.00 (0.00)
kNN	0.85 (0.05)	0.93 (0.02)	0.95 (0.00)	0.95 (0.01)	1.00 (0.00)	1.00 (0.00)
RF	0.88 (0.04)	0.95 (0.00)	0.97 (0.03)	0.97 (0.01)	1.00 (0.00)	1.00 (0.00)
SVM	0.84 (0.04)	0.94 (0.02)	0.95 (0.01)	0.95 (0.01)	1.00 (0.00)	1.00 (0.00)
XGBoost	0.81 (0.04)	0.91 (0.02)	0.95 (0.01)	0.94 (0.01)	0.98 (0.03)	1.00 (0.00)
Kappa						
glmnet	0.48 (0.14)	0.81 (0.05)	0.88 (0.06)	0.86 (0.03)	0.96 (0.06)	1.00 (0.00)
kNN	0.54 (0.11)	0.81 (0.06)	0.86 (0.00)	0.87 (0.02)	1.00 (0.01)	1.00 (0.00)
RF	0.65 (0.13)	0.85 (0.00)	0.93 (0.07)	0.91 (0.02)	1.00 (0.00)	1.00 (0.00)
SVM	0.55 (0.09)	0.82 (0.04)	0.87 (0.02)	0.87 (0.03)	1.00 (0.01)	1.00 (0.00)
XGBoost	0.42 (0.11)	0.74 (0.05)	0.86 (0.03)	0.83 (0.02)	0.95 (0.07)	1.00 (0.00)
AUC						
glmnet						
kNN	0.86 (0.03)	0.98 (0.01)	1.00 (0.00)	0.98 (0.01)	1.00 (0.00)	1.00 (0.00)
RF	0.95 (0.02)	1.00 (0.01)	1.00 (0.00)	0.99 (0.00)	1.00 (0.00)	1.00 (0.00)
SVM	0.92 (0.03)	0.98 (0.01)	1.00 (0.00)	0.99 (0.00)	1.00 (0.00)	1.00 (0.00)
XGBoost	0.94 (0.03)	1.00 (0.01)	1.00 (0.00)	0.99 (0.00)	1.00 (0.00)	1.00 (0.00)

The values in the cells are Mean (SD) of results of each ML algorithm after 10 repetitions of 10-fold cross-validation.
glmnet: lasso and elastic-net regularized generalized linear model; kNN: k-Nearest Neighbors; RF: random forest; SVM: support vector machine; XGBoost: eXtreme Gradient Boosting.

Table 2
Characteristics of important genes as well as expression patterns according to metastatic and non-metastatic samples CRC.

Biomarker	Log FC	adj.P-value	Importance score (scaled score)[a]	Non-metastatic CRC[b] (N = 195)	Metastatic CRC[b] (N = 67)
MMP3	−2.67	3.15E-26	7.39 (1.00)	248.31 (274.52)	31.48 (24.91)
CCDC102B	1.91	4.98E-13	6.56 (0.80)	9.61 (12.74)	30.13 (22.60)
CDH2	1.65	1.29E-12	5.21 (0.49)	17.36 (21.45)	55.42 (61.83)
VWF	−1.45	2.82E-10	4.48 (0.31)	236.03 (156.13)	79.52 (222.18)
MMP1	−2.79	6.15E-12	4.28 (0.27)	291.50 (557.26)	20.10 (97.76)
CYP1B1	1.7	3.25E-11	4.27 (0.26)	27.10 (47.13)	89.30 (112.83)
DIXDC1	−1.34	7.62E-16	4.25 (0.26)	71.21 (45.37)	32.93 (44.20)
LAMC3	1.15	5.64E-09	4.09 (0.22)	19.45 (26.04)	45.54 (44.68)
SERPINA1	1.69	1.14E-11	3.99 (0.20)	581.12 (990.25)	1626.12 (3186.58)
NKX3-2	−1.33	6.63E-12	3.89 (0.17)	15.45 (25.86)	7.45 (4.01)
ALB	1.61	3.48E-11	3.76 (0.14)	9.16 (16.04)	27.84 (68.38)
KRT7	1.13	3.99E-07	3.65 (0.12)	22.77 (19.85)	42.57 (55.35)
TMEM158	−1.59	2.69E-13	3.61 (0.11)	183.11 (153.64)	67.26 (103.20)
GADD45B	1.04	6.26E-09	3.36 (0.05)	10.14 (14.23)	25.79 (39.43)
CYP4B1	1.21	6.95E-07	3.33 (0.04)	3.47 (7.63)	8.99 (22.39)
SCGB1A1	1.03	2.45E-06	3.15 (0.00)	23.06 (19.41)	20.10 (97.76)

CRC: colorectal cancer.
[a] Based on random forest algorithm.
[b] Presented data are median (interquartile range) of gene expression. The median difference of all biomarkers were significant based on Mann-Whitney *U* test (p-value<0.001).

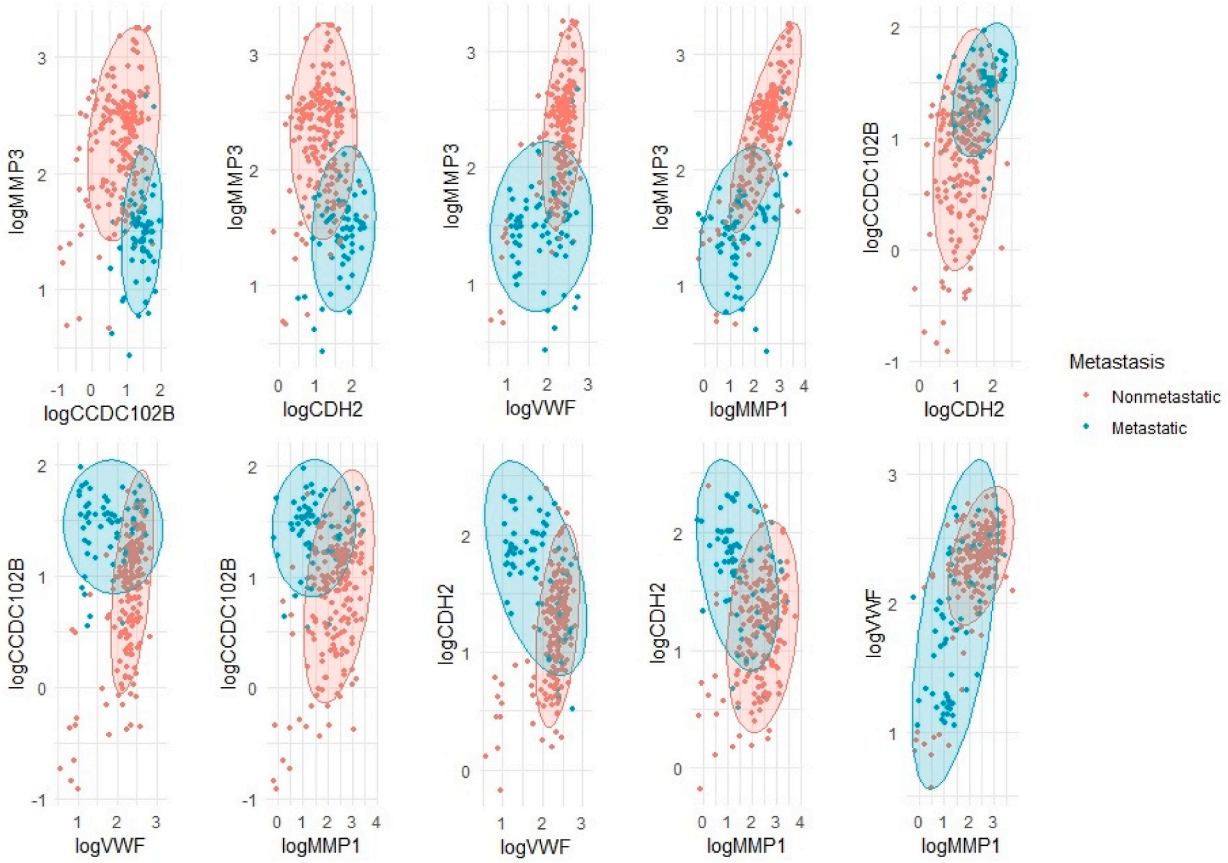


Fig. 2. Discrimination power of top five important genes to differentiate non-metastatic and metastatic colorectal cancer.

of *MMP3* in the metastasis of CRC, some studies have indicated that the expression level of this gene decreased in metastatic tumors compared to primary tumors [35]. For example, Maiti et al. [36], in their study, showed that the expression level of this gene has a reverse correlation with the stage of the tumor.

Results of the current study indicated that *MMP1* same as *MMP3* downregulated in metastatic samples compared with primary

Table 3
Model performance of five machine learning algorithms in the internal validation dataset.

Performance measure	ML algorithm				
	glmnet	kNN	RF	SVM	XGBoost
Accuracy	0.93 (0.03)	0.93 (0.03)	0.97 (0.02)	0.92 (0.02)	0.93 (0.02)
Kappa	0.82 (0.10)	0.81 (0.07)	0.91 (0.05)	0.80 (0.06)	0.81 (0.07)
Specificity	0.96 (0.03)	0.96 (0.02)	0.98 (0.01)	0.96 (0.03)	0.97 (0.02)
Precision	0.88 (0.08)	0.88 (0.05)	0.95 (0.03)	0.89 (0.07)	0.91 (0.06)
Recall	0.85 (0.10)	0.84 (0.08)	0.92 (0.05)	0.81 (0.07)	0.82 (0.09)
F1 score	0.86 (0.07)	0.86 (0.05)	0.93 (0.03)	0.85 (0.04)	0.85 (0.05)

The values in the cells are Mean (SD) of results of each ML algorithm after 10 repetitions of 10 -fold cross-validation.
ML: machine learning; glmnet: lasso and elastic-net regularized generalized linear model; kNN: k-Nearest Neighbors; RF: random forest; SVM: support vector machine; XGBoost: eXtreme Gradient Boosting.

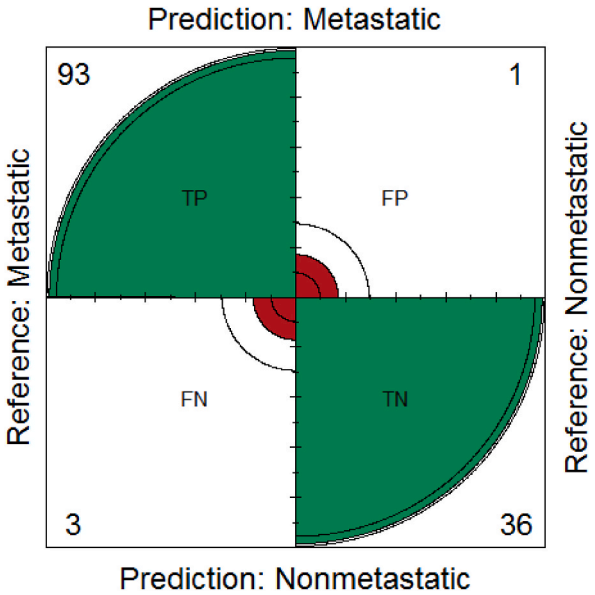


Fig. 3. Performance of the developed biomarker predictive RF model in external and independent dataset. TP: true positive; FP: false positive; FN: false negative; TN: true negative.

tumors. Results of several studies indicated that the upregulation of *MMP1* as another important member of the MMP family has an essential role in the migration, progression, and metastasis of different tumors and is also associated with the prognosis of several cancers [37,38]. Considering these controversial results, the expression level of *MMP1* with metastasis of CRC and the molecular mechanism of its role should be evaluated in biological experiments.

The protein encoded by *CCDC102B*, located on chromosome 18, plays an essential role in centrosome cohesion, and mutations in this gene are common in several tumors [39,40]. Recent studies have indicated that *CCDC102B* plays an essential role in breast cancer metastasis through controlling the RACK1 expression level, NF-κB pathway, and epithelial-mesenchymal transition (EMT) [41–43].

VWF encodes the blood clotting glycoprotein, which plays a role in primary hemostasis after vascular injury. Evaluating *VWF* levels in plasma has indicated that in colorectal patients, the level of this factor increases [44]. On the other hand, *VWF* can the regulate invasion and migration of tumor cells due to its interaction with adhesion molecules of tumor cells such as integrins and extracellular matrix proteins [45]. Recent studies have shown that *VWF* promotes the progression and metastasis of tumors by regulating angiogenesis, vascular permeability, and inflammation [46]. Controversially, Terraube et al. in their study showed that *VWF* suppression is associated with tumor metastasis of melanoma and lung carcinoma [47].

Dysregulation of *CDH2*, also known as N-cadherin, a transmembrane protein, has been observed in several tumors. Several Studies indicated that *CDH2* leads to metastasis by controlling the expression level of MMPs, affecting the EGFR signaling pathway, and controlling the heterotypic cell adhesion [48,49]. This gene plays an essential role in controlling biological processes such as proliferation, metastasis, and response to treatment. Furthermore, the upregulation of CDH2 is related to poor prognosis in different cancers, including CRC [50,51].

Several limitations should be considered when interpreting the results. First, there was an imbalance distribution of the outcome levels in the training dataset that may affect the results. In such situations, classic classifiers (e.g., RF and SVM) tend to be overwhelmed by large classes while ignoring small classes. However, since there was not much difference in the results between the classic classifiers

and optimized distributed gradient boosting (e.g., XGBoost), it seems the class imbalance does not seriously threaten the reliability and robustness of the results. The ability of XGBoost to handle label-imbalanced data has been demonstrated previously [52]. It would have been ideal for the results to be stratified according to the type of metastases; however, in the case of stratification into liver and lung metastases, the imbalance problem would be severe, with 82 % and 92 % of the internal dataset representing the majority class, respectively. In such scenarios, ML algorithms may struggle to identify metastases as minor class data points. Second, although we attempted to determine the external validity of the biomarker predictive model, causal inference and the predictive power of the developed model still require further extensive external validations. Third, it is expected that predictions would be influenced by factors such as age, sex, T stage, N stage, family history of cancer, etc. However, factors affecting metastasis in patients with CRC were not available for inclusion in the analysis. Finally, our work has an exploratory manner that dives deep into identifying a panel of biomarkers through stringent thresholds that may affect the results. For example, the performance of ML algorithms can be a function of the type of hyperparameters for tuning and the search space for hyperparameters. On the other hand, instead of using a narrow-down approach to identify a small subset of biomarkers, a comprehensive approach like a systematic review can be used to identify biomarkers that have already been shown to play a role in predicting metastasis, and then compare their importance in predicting metastasis using databases like GEO database.

In conclusion, we developed a multi-biomarker model for the prediction of CRC metastasis. The ML algorithms, particularly the RF model, demonstrated accurate predictions of metastasis in CRC patients. The decreased expression of *MMP3*, *VWF*, and *MMP1*, coupled with the increased expression of *CDH2* and *CCDC102B*, may play a pivotal role in the prediction of metastasis among CRC patients. More wet-lab experiments and large-scale external validation studies still are needed to confirm the role of the biomarkers including *MMP3, CCDC102B, VWF, MMP1*, and *CDH2* for predicting metastasis among CRC patients.

CRediT authorship contribution statement

Erfan Ayubi: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Sajjad Farashi:** Writing – review & editing, Software, Methodology, Formal analysis, Data curation. **Leili Tapak:** Visualization, Validation, Software, Data curation. **Saeid Afshar:** Writing – review & editing, Supervision, Resources, Data curation.

Ethical approval

The study was reviewed and approved by the ethics committee of Hamadan University of Medical Sciences, Hamadan, Iran (Ethical code: IR.UMSHA.REC.1402.638).

Data and code availability

The data used for current study are all publicly available at https://www.ncbi.nlm.nih.gov/geo/under the accession numbers GSE68468 and GSE41568.

R Codes are available at supplementary file Source Code.

Funding

This study was supported and funded by the Hamadan University of Medical Sciences (Research ID: 140210129085)

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Abbreviation

CRC	Colorectal cancer
GEO	Gene Expression Omnibus
ML	Machine learning
kNN	k-Nearest neighbors
SVM	Support vector machine
RF	Random forest
GBM	gradient boosting machine
XGBoost	eXtreme Gradient Boosting
AUC	area under the curve
DEGs	Differentially expressed genes
RFE	Recursive Feature Elimination
FC	Fold-change

Caret	Classification And REgression Training
CV	cross-validation
MMP3	matrix metalloproteinase-3
CCDC102B	Coiled-Coil Domain Containing 102B
CDH2	cadherin 2
SCGB1A1	secretoglobin family 1A member 1
KRT7	Keratin 7
CYP1B1	Cytochrome P450 1B1
LAMC3	Laminin Subunit Gamma 3
ALB	Albumin
MMP1	Matrix Metallopeptidase-1
DIXDC1	DIX Domain Containing 1
VWF	Von Willebrand Factor
CYP4B1	Cytochrome P450 4B1
NKX3-2	NK3 Homeobox 2
TMEM158	Transmembrane Protein 158
SERPINA1	serpin family A member 1
GADD45B	Growth Arrest And DNA Damage Inducible Beta

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e41443.

References

[1] E. Morgan, M. Arnold, A. Gini, V. Lorenzoni, C.J. Cabasag, M. Laversanne, J. Vignat, J. Ferlay, N. Murphy, F. Bray, Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN, Gut 72 (2023) 338–344, https://doi.org/10.1136/gutjnl-2022-327736.

[2] H. Rumpold, D. Niedersüß-Beke, C. Heiler, D. Falch, H.V. Wundsam, S. Metz-Gercek, G. Piringer, J. Thaler, Prediction of mortality in metastatic colorectal cancer in a real-life population: a multicenter explorative analysis, BMC Cancer 20 (2020) 1149, https://doi.org/10.1186/s12885-020-07656-w.

[3] C. Hackl, P. Neumann, M. Gerken, M. Loss, M. Klinkhammer-Schalke, H.J. Schlitt, Treatment of colorectal liver metastases in Germany: a ten-year population-based analysis of 5772 cases of primary colorectal adenocarcinoma, BMC Cancer 14 (2014) 810, https://doi.org/10.1186/1471-2407-14-810.

[4] J.S. Abelson, H.L. Yeo, J. Mao, J.W. Milsom, A. Sedrakyan, Long-term postprocedural outcomes of palliative emergency stenting vs stoma in malignant large-bowel obstruction, JAMA surgery 152 (2017) 429–435, https://doi.org/10.1001/jamasurg.2016.5043.

[5] L.G. van der Geest, J. Lam-Boer, M. Koopman, C. Verhoef, M.A. Elferink, J.H. de Wilt, Nationwide trends in incidence, treatment and survival of colorectal cancer patients with synchronous metastases, Clin. Exp. Metastasis 32 (2015) 457–465, https://doi.org/10.1007/s10585-015-9719-0.

[6] C. Peixoto, M.B. Lopes, M. Martins, S. Casimiro, D. Sobral, A.R. Grosso, C. Abreu, D. Macedo, A.L. Costa, H. Pais, C. Alvim, A. Mansinho, P. Filipe, P.M.D. Costa, A. Fernandes, P. Borralho, C. Ferreira, J. Malaquias, A. Quintela, S. Kaplan, M. Golkaram, M. Salmans, N. Khan, R. Vijayaraghavan, S. Zhang, T. Pawlowski, J. Godsey, A. So, L. Liu, L. Costa, S. Vinga, Identification of biomarkers predictive of metastasis development in early-stage colorectal cancer using network-based regularization, BMC Bioinf. 24 (2023) 17, https://doi.org/10.1186/s12859-022-05104-z.

[7] W. Xu, Y. He, Y. Wang, X. Li, J. Young, J.P.A. Ioannidis, M.G. Dunlop, E. Theodoratou, Risk factors and risk prediction models for colorectal cancer metastasis and recurrence: an umbrella review of systematic reviews and meta-analyses of observational studies, BMC Med. 18 (2020) 172, https://doi.org/10.1186/s12916-020-01618-6.

[8] F. Del Vecchio, V. Mastroiaco, A. Di Marco, C. Compagnoni, D. Capece, F. Zazzeroni, C. Capalbo, E. Alesse, A. Tessitore, Next-generation sequencing: recent applications to the analysis of colorectal cancer, J. Transl. Med. 15 (2017) 246, https://doi.org/10.1186/s12967-017-1353-y.

[9] I.A. Eilertsen, S.H. Moosavi, J.M. Strømme, A. Nesbakken, B. Johannessen, R.A. Lothe, A. Sveen, Technical differences between sequencing and microarray platforms impact transcriptomic subtyping of colorectal cancer, Cancer letters 469 (2020) 246–255, https://doi.org/10.1016/j.canlet.2019.10.040.

[10] R. Boughriba, G. Sahraoui, I. Chaar, M. Weslati, K. Ayed, D. Ounissi, M. Hazgui, S. Bouraoui, A. Gati, Significant association of MCP1 rs1024611 and CCR2 rs1799864 polymorphisms with colorectal cancer and liver metastases susceptibility and aggressiveness: a case-control study, Cytokine 167 (2023) 156193, https://doi.org/10.1016/j.cyto.2023.156193.

[11] E. Salem, A. Keshvari, A. Mahdavinezhad, A.R. Soltanian, M. Saidijam, S. Afshar, Role of EFNA1 SNP (rs12904) in tumorigenesis and metastasis of colorectal cancer: a bioinformatic analysis and hrm SNP genotyping verification, Asian Pac. J. Cancer Prev. APJCP : APJCP. 23 (2022) 3523–3531, https://doi.org/10.31557/apjcp.2022.23.10.3523.

[12] S. Wu, H. Sun, Y. Wang, X. Yang, Q. Meng, H. Yang, H. Zhu, W. Tang, X. Li, M. Aschner, R. Chen, MALAT1 rs664589 polymorphism inhibits binding to miR-194-5p, contributing to colorectal cancer risk, growth, and metastasis, Cancer Res. 79 (2019) 5432–5441, https://doi.org/10.1158/0008-5472.can-19-0773.

[13] Z. Zhang, H. Jia, Y. Wang, B. Du, J. Zhong, Association of MACC1 expression with lymphatic metastasis in colorectal cancer: a nested case-control study, PLoS One 16 (2021) e0255489, https://doi.org/10.1371/journal.pone.0255489.

[14] L. Chen, H. Li, L. Xie, Z. Zuo, L. Tian, C. Liu, X. Guo, Editorial: big data and machine learning in cancer genomics, Front. Genet. 12 (2021) 749584, https://doi.org/10.3389/fgene.2021.749584.

[15] G.P. Dai, L.P. Wang, Y.Q. Wen, X.Q. Ren, S.G. Zuo, Identification of key genes for predicting colorectal cancer prognosis by integrated bioinformatics analysis, Oncol. Lett. 19 (2020) 388–398, https://doi.org/10.3892/ol.2019.11068.

[16] W. Shi, X. Li, X. Su, H. Wen, T. Chen, H. Wu, M. Liu, The role of multiple metabolic genes in predicting the overall survival of colorectal cancer: a study based on TCGA and GEO databases, PLoS One 16 (2021) e0251323, https://doi.org/10.1371/journal.pone.0251323.

[17] Y. Zhuang, H. Wang, D. Jiang, Y. Li, L. Feng, C. Tian, M. Pu, X. Wang, J. Zhang, Y. Hu, P. Liu, Multi gene mutation signatures in colorectal cancer patients: predict for the diagnosis, pathological classification, staging and prognosis, BMC Cancer 21 (2021) 380, https://doi.org/10.1186/s12885-021-08108-9.

[18] S. Ng, S. Masarone, D. Watson, M.R. Barnes, The benefits and pitfalls of machine learning for biomarker discovery, Cell Tissue Res. (2023), https://doi.org/10.1007/s00441-023-03816-z.

[19] L. Buk Cardoso, V. Cunha Parro, S. Verzinhasse Peres, M.P. Curado, G.A. Fernandes, V. Wünsch Filho, T. Natasha Toporcov, Machine learning for predicting survival of colorectal cancer patients, Sci. Rep. 13 (2023) 8874, https://doi.org/10.1038/s41598-023-35649-9.

[20] E. Nemlander, M. Ewing, E. Abedi, J. Hasselström, A. Sjövall, A.C. Carlsson, A. Rosenblad, A machine learning tool for identifying non-metastatic colorectal cancer in primary care, European journal of cancer (Oxford, England : 1990) 182 (2023) 100–106, https://doi.org/10.1016/j.ejca.2023.01.011.

[21] M.H. Osman, R.H. Mohamed, H.M. Sarhan, E.J. Park, S.H. Baik, K.Y. Lee, J. Kang, Machine learning model for predicting postoperative survival of patients with colorectal cancer, Cancer research and treatment 54 (2022) 517–524, https://doi.org/10.4143/crt.2021.206.

[22] Z. Huang, C. Hu, C. Chi, Z. Jiang, Y. Tong, C. Zhao, An artificial intelligence model for predicting 1-year survival of bone metastases in non-small-cell lung cancer patients based on XGBoost algorithm, BioMed Res. Int. (2020) 3462363, https://doi.org/10.1155/2020/3462363, 2020.

[23] Q. Li, H. Yang, P. Wang, X. Liu, K. Lv, M. Ye, XGBoost-based and tumor-immune characterized gene signature for the prediction of metastatic status in breast cancer, J. Transl. Med. 20 (2022) 177, https://doi.org/10.1186/s12967-022-03369-9.

[24] A.Y. Lin, M.S. Chua, Y.L. Choi, W. Yeh, Y.H. Kim, R. Azzi, G.A. Adams, K. Sainani, M. van de Rijn, S.K. So, J.R. Pollack, Comparative profiling of primary colorectal carcinomas and liver metastases identifies LEF1 as a prognostic biomarker, PLoS One 6 (2011) e16636, https://doi.org/10.1371/journal.pone.0016636.

[25] D.M. Mutch, A. Berger, R. Mansourian, A. Rytz, M.A. Roberts, The limit fold change model: a practical approach for selecting differentially expressed genes from microarray data, BMC Bioinf. 3 (2002) 17, https://doi.org/10.1186/1471-2105-3-17.

[26] P. Baldi, A.D. Long, A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes, Bioinformatics 17 (2001) 509–519, https://doi.org/10.1093/bioinformatics/17.6.509.

[27] H. Jeon, S. Oh, Hybrid-recursive feature elimination for efficient feature selection, Appl. Sci. 10 (2020) 3211.

[28] R. Edgar, M. Domrachev, A.E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, Nucleic acids research 30 (2002) 207–210, https://doi.org/10.1093/nar/30.1.207.

[29] E.D. Huckvale, M.W. Hodgman, B.B. Greenwood, D.O. Stucki, K.M. Ward, M.T.W. Ebbert, J.S.K. Kauwe, I. The Alzheimer's Disease Neuroimaging, C. The Alzheimer's Disease Metabolomics, J.B. Miller, Pairwise correlation analysis of the alzheimer's disease neuroimaging initiative (ADNI) dataset reveals significant feature correlation, Genes 12 (2021), https://doi.org/10.3390/genes12111661.

[30] P.M. Granitto, C. Furlanello, F. Biasioli, F. Gasperi, Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products, Chemometr. Intell. Lab. Syst. 83 (2006) 83–90.

[31] W. You, Z. Yang, G. Ji, Feature selection for high-dimensional multi-category data using PLS-based local recursive feature elimination, Expert Syst. Appl. 41 (2014) 1463–1475.

[32] Z. Pezeshkian, S. Nobili, N. Peyravian, B. Shojaee, H. Nazari, H. Soleimani, H. Asadzadeh-Aghdaei, M. Ashrafian Bonab, E. Nazemalhosseini-Mojarad, E. Mini, Insights into the role of matrix metalloproteinases in precancerous conditions and in colorectal cancer, Cancers 13 (2021), https://doi.org/10.3390/cancers13246226.

[33] S.A. Suhaimi, S.C. Chan, R. Rosli, Matrix Metallopeptidase 3 polymorphisms: emerging genetic markers in human breast cancer metastasis, Journal of breast cancer 23 (2020) 1–9, https://doi.org/10.4048/jbc.2020.23.e17.

[34] M. Liang, J. Wang, C. Wu, M. Wu, J. Hu, J. Dai, H. Ruan, S. Xiong, C. Dong, Targeting matrix metalloproteinase MMP3 greatly enhances oncolytic virus mediated tumor therapy, Translational Oncology 14 (2021) 101221, https://doi.org/10.1016/j.tranon.2021.101221.

[35] A. Ahmadieh-Yazdi, A. Mahdavinezhad, L. Tapak, F. Nouri, A. Taherkhani, S. Afshar, Using machine learning approach for screening metastatic biomarkers in colorectal cancer and predictive modeling with experimental validation, Sci. Rep. 13 (2023) 19426, https://doi.org/10.1038/s41598-023-46633-8.

[36] A. Maiti, I. Okano, M. Oshi, M. Okano, W. Tian, T. Kawaguchi, E. Katsuta, K. Takabe, L. Yan, S. Patnaik, N.C. Hait, Altered expression of secreted mediator genes that mediate aggressive breast cancer metastasis to distant organs, Cancers 13 (2021), https://doi.org/10.3390/cancers13112641.

[37] H. Liu, Y. Kato, S.A. Erzinger, G.M. Kiriakova, Y. Qian, D. Palmieri, P.S. Steeg, J.E. Price, The role of MMP-1 in breast cancer growth and metastasis to the brain in a xenograft model, BMC Cancer 12 (2012) 583, https://doi.org/10.1186/1471-2407-12-583.

[38] W. Zhang, X. Huang, R. Huang, H. Zhu, P. Ye, X. Lin, S. Zhang, M. Wu, F. Jiang, MMP1 overexpression promotes cancer progression and associates with poor outcome in head and neck carcinoma, Comput. Math. Methods Med. 2022 (2022) 3058342, https://doi.org/10.1155/2022/3058342.

[39] B.R. Druliner, X. Ruan, H. Sicotte, D. O'Brien, H. Liu, J.A. Kocher, L. Boardman, Early genetic aberrations in patients with sporadic colorectal cancer, Mol. Carcinog. 57 (2018) 114–124, https://doi.org/10.1002/mc.22738.

[40] Y. Xia, N. Huang, Z. Chen, F. Li, G. Fan, D. Ma, J. Chen, J. Teng, CCDC102B functions in centrosome linker assembly and centrosome cohesion, J. Cell Sci. 131 (2018), https://doi.org/10.1242/jcs.222901.

[41] J. Si, R. Guo, B. Xiu, W. Chi, Q. Zhang, J. Hou, Y. Su, J. Chen, J. Xue, Z.M. Shao, J. Wu, Y. Chi, Stabilization of CCDC102B by loss of RACK1 through the CMA pathway promotes breast cancer metastasis via activation of the NF-κB pathway, Frontiers in oncology 12 (2022) 927358, https://doi.org/10.3389/fonc.2022.927358.

[42] J. Si, R. Guo, Abstract 1112: CCDC102B Promotes Metastatic Cascade in Breast Cancer by Activating NF-Kb via Down-Regulating RACK1, 2019.

[43] J. Si, R. Guo, Abstract 1112: CCDC102B promotes metastatic cascade in breast cancer by activating NF-κB via down-regulating RACK1, Cancer Res. 79 (2019), https://doi.org/10.1158/1538-7445.am2019-1112, 1112-1112.

[44] V.S. Schellerer, L. Mueller-Bergh, S. Merkel, R. Zimmermann, D. Weiss, A. Schlabrakowski, E. Naschberger, M. Stürzl, W. Hohenberger, R.S. Croner, The clinical value of von Willebrand factor in colorectal carcinomas, American journal of translational research 3 (2011) 445–453.

[45] V. Terraube, R. Pendu, D. Baruch, M.F.B.G. Gebbink, D. Meyer, P.J. Lenting, C.V. Denis, Increased metastatic potential of tumor cells in von Willebrand factor-deficient mice, J. Thromb. Haemostasis 4 (2006) 519–526, https://doi.org/10.1111/j.1538-7836.2005.01770.x.

[46] S. Patmore, S.P.S. Dhami, J.M. O'Sullivan, Von Willebrand factor and cancer; metastasis and coagulopathies, J. Thromb. Haemostasis 18 (2020) 2444–2456, https://doi.org/10.1111/jth.14976.

[47] V. Terraube, I. Marx, C.V. Denis, Role of von Willebrand factor in tumor metastasis, Thromb. Res. 120 (Suppl 2) (2007) S64–S70, https://doi.org/10.1016/s0049-3848(07)70132-9.

[48] K.M. Mrozik, O.W. Blaschuk, C.M. Cheong, A.C.W. Zannettino, K. Vandyke, N-cadherin in cancer metastasis, its emerging role in haematological malignancies and potential as a therapeutic target in cancer, BMC Cancer 18 (2018) 939, https://doi.org/10.1186/s12885-018-4845-0.

[49] Z.-Q. Cao, Z. Wang, P. Leng, Aberrant N-cadherin expression in cancer, Biomed. Pharmacother. 118 (2019) 109320, https://doi.org/10.1016/j.biopha.2019.109320.

[50] T. Zhang, K. Yuan, Y. Wang, M. Xu, C. Shirong, C. Chen, J. Ma, Identification of candidate biomarkers and prognostic analysis in colorectal cancer liver metastases, Frontiers in oncology 11 (2021) 652354, https://doi.org/10.3389/fonc.2021.652354.

[51] X. Yan, L. Yan, S. Liu, Z. Shan, Y. Tian, Z. Jin, N-cadherin, a novel prognostic biomarker, drives malignant progression of colorectal cancer, Mol. Med. Rep. 12 (2015) 2999–3006, https://doi.org/10.3892/mmr.2015.3687.

[52] C. Wang, C. Deng, S. Wang, Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost, Pattern Recogn. Lett. 136 (2020) 190–197.