Two-phase Classification of Colorectal Cancer Stages

Frosina Stojanovska, Viktorija Velinovska, Monika Simjanoska and Ana Madevska Bogdanova
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of Macedonia
stojanovska.frose@gmail.com, velinovska.viktorija@gmail.com,
monika.simjanoska@finki.ukim.mk, ana.madevska.bogdanova@finki.ukim.mk

***Abstract*—Staging of colorectal cancer is one of the essential factors required to identify the patient's true therapy for recovery. Despite the various clinical colorectal cancer staging methods, this problem remains critical for personalized stage determination. In this paper, we study the problem of colorectal cancer stage determination using gene expression data obtained from DNA microarrays. The goal is to construct a supervised machine learning classification model that will be able to detect the stage of colorectal cancer, that is, the model should be able to separate the stages utilizing 11 biomarkers as features.**

Dataset resampling and analyzing the errors between the real and the predicted class during validation phase, led to the creation of two-phase classification model, dividing the main problem of determining the stage of the colorectal cancer into sub-problems. In the first phase of classification, it is necessary to create a classification model that will successfully divide the data between two groups obtained by joining stage I and IV as one sub-group, and stage II and III as the second sub-group. Once an instance of the data set is classified into one of the combined classes, according to this class, the second level classification reveals the true cancer stage of the instance. Random Forest is the machine learning algorithm that performed best in all the experiments, compared to KNN, SVM, Naive Bayes and MLP.

***Keywords*—gene expression, colorectal cancer, stage detection, machine learning, ensemble methods**

I. Introduction

Colorectal cancer (CRC) is malignant cancer located in the colon and/or rectum. According to the statistics from the World Health Organization (WHO) [1] provided in 2017, the cancer is one of the leading sources of death worldwide, and colorectal cancer is the third most common type of cancer that occurs in men and women.

After the cancer is diagnosed, it is essential to find the level of cancer expansion in the affected body part. This is the process of cancer staging, which helps the doctors to choose the most appropriate treatment for the recovery of the patient. There are four stages of colorectal cancer spread in the AJCC TNM (Tumor size, Lymph Nodes affected, Metastases) staging system [2], starting from stage I (1) to stage IV (4), and additionally stage 0 representing a very early phase of cancer. With this order, an earlier stage refers to a lower degree of cancer. Histopathology is used in clinical practice for discovering the cancer stage, with analysis of the local tumor invasion and the presence of CRC metastases in lymph nodes. Histologic staging has difficulty detecting the cancer stage in individuals, so there is a need for more sensitive and better methods [3].

Cancer is a disease caused by several genetic and epigenetic alterations [4]. These genetic changes can lead to an unusual growth of the cells that are transforming into cancer cells. Cancer research includes solutions from bioinformatics, for instance, diagnostic protocols or pattern discovery in cancer by analyzing biological data, especially of the omics data [5]. The progression of omics data analysis with bioinformatics technologies involves the integration of huge amount of data, including genomics, transcriptomics and proteomics data from many different sources. The multi-omics analysis is continuously more popular in biomedical research, and as a result authors in [6] had built a freely available platform LinkedOmics for analysis and comparison of cancer multi-omics data within and across multiple cancer types.

Machine learning methods are rising as a solution to many problems in distinct domains. These techniques are utilized to model the progression and treatment of cancerous conditions [7]. Machine learning, with its supervised, semi-supervised, unsupervised, or even reinforcement learning methods, has the ability to give an interpretation, or, a possible solution of many biological problems [8]–[10]. The authors in [11] give an overview of various machine learning models applied to cancer prognosis and prediction.

Some implemented machine learning algorithms rely on gene biomarkers as features to build the models. Biomarkers have a key role in cancer disease discovery, treatment selection, drug discovery, and personalized medicine [12]. Although there are plenty of studies that report biomarkers as significant related to some disease, there are still very few of them validated of proven and robust clinical utility [13].

Lately, many researchers study and analyze the gene expression profile data associated with CRC. The authors in [14] inferred a colon cancer gene regulatory network and studied its functional and structural meaning using gene expression data. The goal in this direction is to make a comparative analysis of this kind of networks of more than one cancer networks [14]. The research in [15] introduces a study for finding the potential key candidate genes and pathways in CRC from the differentially expressed genes (DEGs).

In this paper, we investigate and present a solution to the

problem of detecting the CRC stage employing supervised machine learning models built with gene expression data in infected cells. The rest of the paper is organized as follows. Section II presents the dataset and the methods used to obtain this dataset. Additionally, it gives details of the machine learning methods used to build the model of stage classification. The details of the experiments and the results are given in Section III. Finally, the last section, Section IV, recapitulates the main findings and offers suggestions for future work.

II. Materials and Methods

DNA microarrays are used to study the extent to which certain genes are active in cells and tissues. Two widely used types of DNA microarrays are Affymetrix and Illumina chips [16]. More detailed information about this technology is described in [17].

A. Colorectal Cancer Dataset

The dataset used in this paper consists of 657 instances with features comprised of gene expression from 11 genes, selected to be the biomarkers associated with CRC, and 1 feature for the CRC stage. The distribution of the four CRC stages is given as:

- *Stage I* - gene expressions from 137 patients.
- *Stage II* - gene expressions from 257 patients.
- *Stage III* - gene expressions from 182 patients.
- *Stage IV* - gene expressions from 81 patients.

The dataset is constructed by merging several CRC datasets from Gene Expression Omnibus database [18] (whose identifiers are provided in [19]). The 11 biomarkers extracted by the analysis presented in [19] showed influence in colorectal cancer determination. These biomarkers are used in this paper as features to investigate their importance as gene biomarkers in colorectal cancer stage determination.

The selected biomarkers genes are:

- *CDH3* - Cadherin 3 (or P-cadherin) is a protein-coding gene that encodes a classical cadherin of the cadherin superfamily. This gene is located on the chromosome 16 and is associated with specific hereditary genetic disorders and several cancers including CRC [20].
- *CHGA* - Chromogranin A is the gene that encodes a protein that is part of the granin family of neuroendocrine secretory proteins. CHGA is used as an indicator of neuroendocrine tumors including carcinoids [21], [22].
- *DHRS9* - Dehydrogenase/reductase 9 is the official name of this gene that encodes a protein which has an oxidoreductase activity toward hydroxysteroids and is a member of short-chain dehydrogenases/reductases (SDR) family. Paper [23] provides evidence for association of the decreased expression of DHRS9 with disease progression and poor outcome of CRC patients.
- *GUCA2A* - GUCA2A or guanylate cyclase activator 2A is an endogenous activator of intestinal guanylate cyclase. The differential expression of this gene in CRC was associated with the tumor stage in [24].
- *GUCA2B* - Guanylate cyclase activator 2B encodes a preproprotein that is proteolytically processed to generate multiple protein products from the guanylin family. This gene was one of the six colorectal cancer related genes in [25].
- *HPGD* - Hydroxyprostaglandin dehydrogenase 15-(NAD) encodes nonmetalloenzyme alcohol dehydrogenase protein responsible for the metabolism of prostaglandins.
- *MMP3* - Matrix metallopeptidase 3, as a gene from the cluster of MMP genes, encodes protein from the matrix metalloproteinase (MMP) family. In [26] MMP3 is introduces as a prognostic factor of tumor progression in three common poor prognosis tumor types (pancreatic, pulmonary, and mammary carcinoma).
- *MMP7* - Matrix metallopeptidase 7 is another gene from the cluster of MMP genes. This gene is overexpressed in association with CRC liver metastases in paper [27].
- *PYY* - Peptide YY is the full name of the gene that encodes preprotein as one of the neuropeptide Y (NPY) family of peptides.
- *SCG2* - Secretogranin II encodes one type of neuroendocrine secretory proteins. This gene was overexpressed in advanced prostate cancer as shown in paper [28].
- *VIP* - Vasoactive intestinal peptide (or VIP) encodes a glucagon protein. This gene is expressed in several tissues, most abundant in pancreatic islets cells and nerve ganglion [29].

The biomarkers importance has been experimentally investigated by using the Random Forest method. Fig. 1 presents the importance of the biomarkers in descending order, showing the gene MMP7 to be the most important and the gene HPGD to be the least important biomarker for separation of the stages with the Random Forest model.

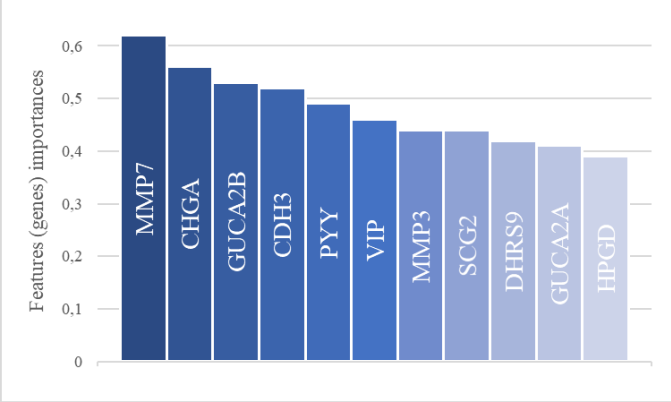

Fig. 1. Feature importance of the biomarkers according to Random Forest method.

B. The Methodology

Several supervised machine learning algorithms were built to classify the stage of expansion of the CRC. The models were supposed to find the separation of the CRC stages using the dataset explained in the previous Section II-A.

This section provides an overview of the selected and applied algorithms: Support Vector Machines, K Nearest Neighbors, Multilayer Perceptron, Naive Bayes and Random Forest.

1) SVM: Support vector machines (SVM) are the standard machine learning technique utilized for many problems. SVMs take the data as input and process it into a large dimensional space. Although SVM can be quite complex, considering the small dimension of our dataset used for training, instance and feature size, this was not a problem in our case. The SVM classifier is important with our approach not only because it promises a good performance, as shown with many other similar implementations, but also it is a model that can capture the multivariate statistical properties of our data.

We need a model that distinguishes four different CRC stages, although SVM works with a binary class. Consequently, we applied SVM with a pairwise classification (one vs one). The proper kernel function holds the ability to model the high-dimensional associations from the data. We have selected the following options for the kernel function: polynomial kernel, Pearson VII function-based universal kernel (PUK), and radial basis function kernel (RBF kernel). PUK kernel function, shown in (1), had the best performance from all kernel functions, so the results in Section III refer to SVM model with PUK kernel function.

$$K(x_i, x_j) = \frac{1}{[1 + (2\sqrt{\| x_i - x_j \|^2}\sqrt{2^{1/\omega} - 1/\sigma})^2]^{\omega}} \quad (1)$$

2) KNN: We used the IBk algorithm to implement the KNN classification. This algorithm actually represents the KNN algorithm, where IB refers to instance-based (the other name under which the nearest neighbors are known), while k allows us to specify the number of closest neighbors. KNN as a lazy approach works without creating a model and classifies a new data point with the data itself. We can notice that this method is much simpler than SVM. KNN is not limited to linearity, so it can capture even nonlinear relations between the features. This factor made this algorithm relevant to our problem regarding that we did not know the type of interaction between the genes.

We use the Euclidean distance as a measure of calculating the closeness of data points, having in mind that all the features are actually real numbers. To determine the optimal value for the parameter k, we considered a space of values ranging from 3 to 21. The optimal k-value in most of the experiments was 15. Also, the best results required standardization of the attributes, that is, the gene expression of the 11 biomarkers. Apart from the distance measure and the number of closest neighbours, KNN has another parameter - vote weighting. Initially, the weights of every data point were equal. Setting the weight to be the inverse distance $(1/distance)$ the method remarkably improved its performance.

3) MLP: In the last few years, Multilayer Perceptron (MLP) has become one of the most promising methods in machine learning. This neural network consists of hidden layers with multiple perceptrons that enable the modelling of any function required for achieving the best separation of the data instances. With this property, MLP was one of the algorithms selected to model the function of the dataset with unknown feature relations.

We implemented the network with one hidden layer. Adding additional layers did not bring any gain, which is expected given the small size of the dataset. The backpropagation algorithm was used to change the weights of the neurons in the process of training. These weights were adjusted using the gradient descent and squared error loss function. In this study, we use the sigmoid function as the activation function.

4) Naive Bayes: The classification with the Naive Bayes model is based on the Bayesian theorem which uses independent assumptions between the predicates. This classifier is easy to build, so it is also suitable for large datasets. This algorithm is a simple technique for constructing a classifier, where the model in this study is built with probabilities obtained through the features of gene expressions and CRC stages in the dataset. We do not have information about the independence of the gene biomarkers. However, despite the naive design and obviously too simple assumptions, this classifier has proven to work well in very complex real situations, overcoming other much more complex classifiers, so it is part of our experiments.

5) Random Forest: Random Forest is an ensemble learning procedure called Bootstrap Aggregation, or, Bagging, adopted for classification, regression and similar problems [30]. This method has excellent performance in classification tasks, equivalent to standard methods as SVMs. Random Forest has promising features including the ability for classification of both two-class and multi-class problems of more than two classes. Also, it is able to measure the feature (gene) importance. Another advantage is that the parameter fine-tuning is simple, with a selection of small numbers of parameters including the number of input features, the number of trees in each forest, as well as the minimum size of the leaf nodes.

The class determination, that is, the classification, is obtained by the mean of classes received by all trees. With this, Random Forest tries to fix the overfitting that trees do with the training datasets. This appeared as a method that would help to find the perfect decision boundaries between CRC stages. This algorithm works efficiently with both large and small datasets. It can handle a huge number of input features without removing some of them.

Generated "forests" can be stored for the next use of other datasets. Also, it is able to calculate closeness between pairs of cases that can be used for clustering, finding outliers, or to give interesting views of the data (by scaling), which can be used to visualize the correlations in our dataset.

III. Experiments and Results

To build and evaluate the described models in the previous section, we used the Weka software tool [31], as well as the web tool ArrayMinning described in [32]. We performed the experiments according to the complexity of the method, starting from Naive Bayes, up to MLP and SVM.

All the classification models were tested using cross-validation with k folds, where for the parameter k, we assigned

a value of 10. The results obtained with the use of all the classifiers are shown in Table I. This validation technique is used in all of the experiments in this section.

The performance of the techniques has been measured by using *correctly classified (CC) instances*, *Area Under the Curve (AUC)*, *Kappa statistics (KS)* and *Mean Absolute Error (MAE)*. We use CC to show the overall accuracy of the classification of the method, and AUC refers to weighted average of the area under the Receiver Operating Characteristic (ROC) curve of every class. KS denotes the reliability of the method, i.e. measures how the improvement of the predictor is relative to a random predictor (1 means perfect predictor, 0 means the predictor is no better than a random one). The KS metrics is given as

$$KS = \frac{p_a - p_r}{1 - p_r} \tag{2}$$

where p_a is the success rate of the actual predictor and p_r is the success rate of a random predictor. MAE is the sum of the absolute differences between predictions and actual values. It is commonly used in regression models, and for the classification is defined as

$$MAE = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k} \mid a_j - p_j \mid}{n} \tag{3}$$

where n is the number of instances in the test set, k refers to the number of classes, a_j is the actual class value (1 if the particular instance class is j and 0 otherwise), and p_j is the predicted probability of the model for the instance to be classified as class j. We compared the performance of the methods using the CC and AUC metrics and used the other metrics as a profound observation of the precision and separation ability of the models.

A. One-phase Classification

The performance of the models shown in Table I is not satisfactory, i.e., the models are not capable of separating the stages of cancer. Therefore, we tried to pre-process the dataset with different methods before training the models. Of all the pre-processing techniques that were performed, the only thing that made a significant improvement was the sampling of the dataset. This method modified the dataset while leaving the same number of instances. In fact, samples were randomly selected from the "old" dataset in order to build the "new" dataset, preserving the initial distribution of the classes. When selecting a sample of the dataset, it can be re-selected as a sample in the next iterations - the process called sampling with replacement. The last iteration is actually the iteration with which the dataset has the same number of samples as in the start before pre-processing.

Table II holds the results of the classification models after performing sampling with replacement. From the results in this table, Random Forest can be distinguished as the best model for classification, which in the previous attempt, shown in Table I, performs slightly less than SVM and KNN. KNN is very close to the performance of Random Forest. Next are the models of SVM and MLP, while Naive Bayes model is the weakest model in this combination of classification of the given data set.

TABLE I
CLASSIFICATION RESULTS

Metric	SVM	KNN (k=15)	Multilayer Perceptron	Naive Bayes	Random Forest
CC	**44.44%**	**44.44%**	41.25%	42.77%	42.77%
AUC	0.594	0.646	0.624	0.654	**0.658**
KS	0.141	**0.180**	0.114	0.178	0.156
MAE	0.327	0.323	**0.313**	**0.313**	0.320

TABLE II
CLASSIFICATION RESULTS AFTER DATASET SAMPLING

Metric	SVM	KNN (k=15)	Multilayer Perceptron	Naive Bayes	Random Forest
CC	57.08%	73.67%	55.40%	43.99%	**75.95%**
AUC	0.699	0.921	0.735	0.670	**0.926**
KS	0.352	0.624	0.370	0.206	**0.656**
MAE	0.311	**0.160**	0.249	0.310	0.187

Fig. 2 shows the difference in the performance of the models, i.e. the improvement of the accuracy or the percentage of correctly classified instances with the original dataset before sampling, and the dataset after sampling.

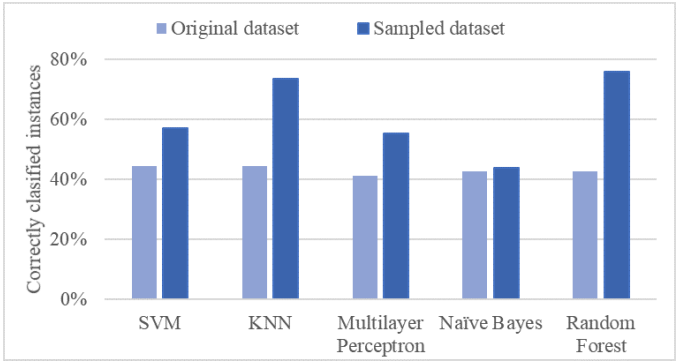

Fig. 2. Difference in the accuracy of the models before and after dataset sampling.

B. Two-phase Classification

Analyzing the errors between the real and predicted classes during testing, we observed an association between the first and fourth stage. With this property, we decided to create a two-phase classification model, where the main problem of determining the stage of the cancer is divided into solving two sub-problems. Therefore, we created two sub-groups of the cancer stages from the resampled dataset, where the first sub-group combines the first and the fourth stage and the second sub-group is a compound of the second and third

cancer stage. Fig. 3 provides a visual representation of the two-phase classification model of CRC stages.

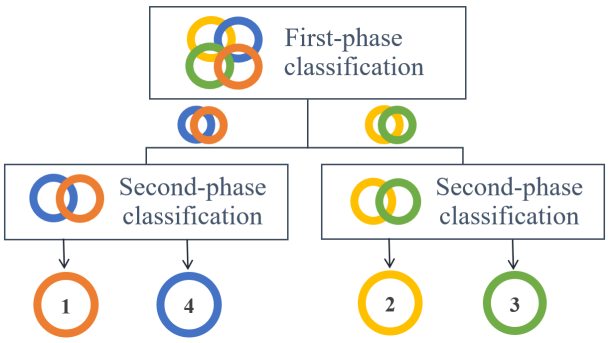

Fig. 3. Visual representation of the two-phase classification model - The first phase is the detection of the sub-groups, and the second phase separates the sub-groups into the real (actual) cancer stages. Cancer stage I is the orange circle, stage II is represented by the yellow circle, stage III is the green one, and the blue circle is stage IV.

The first stage of the classification model considered the separation of the instances from the sub-groups. This stage uses the same machine learning models as in the previous problem. The results are presented in Table III. Random Forest, obtained the highest accuracy of 87%, that is, this algorithm can classify the instances of two sub-groups, with reasonably high correctness. The other algorithms are not nearly satisfactory as Random Forest, especially Naive Bayes.

TABLE III
FIRST-STAGE CLASSIFICATION RESULTS IN THE TWO-PHASE CLASSIFICATION MODEL

Metric	SVM	KNN (k=15)	Multilayer Perceptron	Naive Bayes	Random Forest
CC	76.10%	82.57%	75.34%	68.49%	**87.21%**
AUC	0.663	0.935	0.766	0.699	**0.946**
KS	0.377	0.615	0.459	0.268	**0.700**
MAE	0.239	**0.186**	0.278	0.371	0.211

After the first-stage classification determines the aggregate class, i.e. the sub-group of the instance, depending on the identified sub-group, the exact stage of cancer should be determined. Hence, the next sub-problem required finding two separate classifiers. The first classification model knows how to divide the first sub-group into the first or fourth stage of cancer. Respectively, the second model splits the second subgroup into the second and third stage of cancer. Again, the models were built with the same machine learning techniques.

The results of the classification model that distinguish the first and fourth CRC stage are given in Table IV. This classification divides the first sub-group. Random Forest again shows the best performance with an accuracy of 85.32%. Table V presents the results of the other classification model in the second-stage that separates the second sub-group into the second and third CRC stage. As before, Random Forest is again dominating in the process of separation of the classes,

TABLE IV
SECOND-STAGE CLASSIFICATION RESULTS FOR THE FIRST SUB-GROUP

Metric	SVM	KNN (k=11)	Multilayer Perceptron	Naive Bayes	Random Forest
CC	77.06%	83.03%	79.36%	62.84%	**85.32%**
AUC	0.707	0.946	0.758	0.666	**0.956**
KS	0.472	0.628	0.538	0.248	**0.678**
MAE	0.223	0.180	0.115	0.200	**0.106**

with an accuracy of 83.37%. In the second-stage classification, KNN, MLP and SVM have notable outcomes, whereas Naive Bayes is not appropriate for resolving this difficulty.

TABLE V
SECOND-STAGE CLASSIFICATION RESULTS FOR THE SECOND SUB-GROUP

Metric	SVM	KNN (k=9)	Multilayer Perceptron	Naive Bayes	Random Forest
CC	73.80%	79.73%	78.36%	60.59%	**83.37%**
AUC	0.716	0.916	0.746	0.645	**0.923**
KS	0.442	0.578	0.553	0.137	**0.653**
MAE	0.226	0.210	0.220	0.212	**0.128**

Fig. 4 illustrates the two-phase classification model, with accuracy performance of each particular classification model for every sub-task.

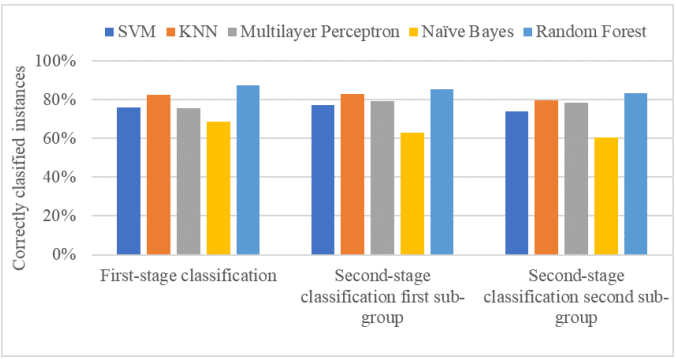

Fig. 4. Performance of the models in the individual tasks in the two-phase classification model.

IV. CONCLUSIONS

By using machine learning techniques, we classify four stages of colorectal cancer in patients, using the gene expression of 11 biomarkers obtained by DNA microarray technology. Recent research has shown that changes in gene expression are associated with different types of cancer.

The choice of best machine learning algorithm to be applied, is based on the nature of the problem and the data set that is used. We used several methods for building the classification model: KNN, SVM, MLP, Naive Bayes and Random Forest. With the initial set, we did not obtain good results, however, when a dataset resampling is applied, the classification significantly improved. The model with Random Forest stands out

as the best classifier model with an accuracy of 76%, along with KNN with 74% accuracy, which was not as satisfactory as we expected.

Considering the unexpected association between the first and the fourth stage, a two-phase classification model was created. In the first phase, the model divides the data between two sub-groups obtained by joining the first and fourth stage as one sub-group, and the second and third stage as the second sub-group. As the best classifier for this case, Random Forest stands out with an accuracy of 87%. The second phase contains two classifiers to divide each individual subgroup to obtain the right cancer stage of the instance. Random Forest again shows the best performance - for the classification of the first and fourth stage the accuracy is 85%, while for the classification of the second and third stage of cancer the accuracy is 83%.

Given the results, we can conclude that the ensemble machine learning methods, represented by Random Forest, along with slightly worse KNN, provide better modelling of the CRC biomarkers gene expressions. The importance of the developed two-phase classification of gene expression for other cancers or other biomarkers remains to be revealed.

The future work will include a deeper analysis of the problem and the CRC data. One option to consider is transfer learning which has the potential of combining previously gained knowledge and solve new related issues. It is mostly implemented with Deep Learning architectures, however, Random Forests are also capable of model transferring. This can be very helpful with very small and limited datasets, as in our case.

REFERENCES

[1] (2017) World health organization. [Online]. Available: http://www.who.int/en/

[2] S. B. Edge and C. C. Compton, "The american joint committee on cancer: the 7th edition of the ajcc cancer staging manual and the future of tnm," *Annals of surgical oncology*, vol. 17, no. 6, pp. 1471–1474, 2010.

[3] S. A. Bustin and S. Dorudi, "Gene expression profiling for molecular staging and prognosis prediction in colorectal cancer," *Expert review of molecular diagnostics*, vol. 4, no. 5, pp. 599–607, 2004.

[4] A. Balmain, J. Gray, and B. Ponder, "The genetics and genomics of cancer," *Nature genetics*, vol. 33, p. 238, 2003.

[5] B. Stransky and P. Galante, "Application of bioinformatics in cancer research," in *An Omics Perspective on Cancer Research*. Springer, 2010, pp. 211–233.

[6] S. V. Vasaikar, P. Straub, J. Wang, and B. Zhang, "Linkedomics: analyzing multi-omics data within and across 32 cancer types," *Nucleic acids research*, vol. 46, no. D1, pp. D956–D963, 2017.

[7] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and structural biotechnology journal*, vol. 13, pp. 8–17, 2015.

[8] P. Baldi and S. Brunak, *Bioinformatics: the machine learning approach*. MIT press, 2001.

[9] I. Inza, B. Calvo, R. Armañanzas, E. Bengoetxea, P. Larrañaga, and J. A. Lozano, "Machine learning: an indispensable tool in bioinformatics," in *Bioinformatics methods in clinical research*. Springer, 2010, pp. 25–48.

[10] P. Larranaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armananzas, G. Santafé, A. Pérez *et al.*, "Machine learning in bioinformatics," *Briefings in bioinformatics*, pp. 86–112, 2006.

[11] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer informatics*, vol. 2, p. 117693510600200030, 2006.

[12] P. Carrigan and T. Krahn, "Impact of biomarkers on personalized medicine," in *New Approaches to Drug Discovery*. Springer, 2015, pp. 285–311.

[13] G. Novelli, C. Ciccacci, P. Borgiani, M. P. Amati, and E. Abadie, "Genetic tests and genomic biomarkers: regulation, qualification and validation," *Clinical cases in mineral and bone metabolism*, vol. 5, no. 2, p. 149, 2008.

[14] F. Emmert-Streib, R. de Matos Simoes, G. Glazko, S. McDade, B. Haibe-Kains, A. Holzinger, M. Dehmer, and F. C. Campbell, "Functional and genetic analysis of the colon cancer network," *BMC bioinformatics*, vol. 15, no. 6, p. S6, 2014.

[15] Y. Guo, Y. Bao, M. Ma, and W. Yang, "Identification of key candidate genes and pathways in colorectal cancer by integrated bioinformatical analysis," *International journal of molecular sciences*, vol. 18, no. 4, p. 722, 2017.

[16] M. Barnes, J. Freudenberg, S. Thompson, B. Aronow, and P. Pavlidis, "Experimental comparison and cross-validation of the affymetrix and illumina gene expression analysis platforms," *Nucleic acids research*, vol. 33, no. 18, pp. 5914–5923, 2005.

[17] V. Trevino, F. Falciani, and H. A. Barrera-Saldaña, "Dna microarrays: a powerful genomic tool for biomedical and clinical research," *Molecular Medicine*, vol. 13, no. 9-10, p. 527, 2007.

[18] (2013) Gene expression omnibus. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/

[19] M. Simjanoska, A. M. Bogdanova, and S. Panov, "Gene ontology analysis of colorectal cancer biomarkers probed with affymetrix and illumina microarrays." in *IJCCI*, 2013, pp. 396–406.

[20] A. F. Vieira and J. Paredes, "P-cadherin and the journey to cancer metastasis," *Molecular cancer*, vol. 14, no. 1, p. 178, 2015.

[21] K. A. Mirkin, C. S. Hollenbeak, and J. Wong, "Impact of chromogranin a, differentiation, and mitoses in nonfunctional pancreatic neuroendocrine tumors 2 cm," *journal of surgical research*, vol. 211, pp. 206–214, 2017.

[22] W. Rogowski, E. Wachuła, A. Lewczuk, A. Kolesińska-Ćwikła, E. Iżycka-Świeszewska, V. Sulżyc-Bielicka, and J. B. Ćwikła, "Baseline chromogranin a and its dynamics are prognostic markers in gastroenteropancreatic neuroendocrine tumors," *Future Oncology*, vol. 13, no. 12, pp. 1069–1079, 2017.

[23] L. Hu, H.-Y. Chen, T. Han, G.-Z. Yang, D. Feng, C.-Y. Qi, H. Gong, Y.-X. Zhai, Q.-P. Cai, and C.-F. Gao, "Downregulation of dhrs9 expression in colorectal cancer tissues and its prognostic significance," *Tumor Biology*, vol. 37, no. 1, pp. 837–845, 2016.

[24] Y. Chen, Y. Zhu, H. Feng, Y. Liu, J. Qian, Y. Fan, and D. Li, "Differential expression of guanylin in colorectal cancer," *Zhonghua wei chang wai ke za zhi= Chinese journal of gastrointestinal surgery*, vol. 12, no. 5, pp. 515–517, 2009.

[25] B.-Q. Li, T. Huang, L. Liu, Y.-D. Cai, and K.-C. Chou, "Identification of colorectal cancer related genes with mrmr and shortest path in protein-protein interaction network," *PloS one*, vol. 7, no. 4, p. e33393, 2012.

[26] C. Mehner, E. Miller, A. Nassar, W. R. Bamlet, E. S. Radisky, and D. C. Radisky, "Tumor cell expression of mmp3 as a prognostic factor for poor survival in pancreatic, pulmonary, and mammary carcinoma," *Genes & cancer*, vol. 6, no. 11-12, p. 480, 2015.

[27] Z.-S. Zeng, W.-P. Shu, A. M. Cohen, and J. G. Guillem, "Matrix metalloproteinase-7 expression in colorectal cancer liver metastases: evidence for involvement of mmp-7 activation in human cancer metastases," *Clinical Cancer Research*, vol. 8, no. 1, pp. 144–148, 2002.

[28] M. Courel, F.-Z. El Yamani, D. Alexandre, H. El Fatemi, C. Delestre, M. Montero-Hadjadje, F. Tazi, A. Amarti, R. Magoul, N. Chartrel *et al.*, "Secretogranin ii is overexpressed in advanced prostate cancer and promotes the neuroendocrine differentiation of prostate cancer cells," *European Journal of Cancer*, vol. 50, no. 17, pp. 3039–3049, 2014.

[29] I. Gozes, M. Bodner, Y. Shani, and M. Fridkin, "Structure and expression of the vasoactive intestinal peptide (vip) gene in a human tumor," *Peptides*, vol. 7, pp. 1–6, 1986.

[30] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. CRC press, 2012.

[31] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[32] E. Glaab, J. M. Garibaldi, and N. Krasnogor, "Arraymining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization," *BMC bioinformatics*, vol. 10, no. 1, p. 358, 2009.