Improved Multiple Part Algorithm (IMPA) to extract multiple solutions for RNA sequence classification problem

Naoual Guannoni*, Faouzi Mhamdi†, Mourad Elloumi‡
*†‡ Laboratory of Technologies of Information and Communication and Electrical Engineering (LaTICE),
National Higher School of Engineers of Tunis (ENSIT), University of Tunis, Tunisia
*Faculty of Science of Tunis, University of Tunis el Manar, Tunis, Tunisia
† Higher Institute of Applied Languages and Computer Science of Beja,University of Jendouba, Tunisia
*nawel.gannouni90@gmail.com, †faouzi.mhamdi@ensi.rnu.tn, ‡Mourad.Elloumi@gmail.com

***Abstract*—The methods, which extract knowledge from Next Generation Sequencing Data (NGS) are highly requested nowadays. The attention to analysis biomedical data is increasing proportionally. In this work, we focus to elicit and discovery a higher amount of knowledge by computing many classification models in a single run, and therefore to identify most of the features related to an investigated class. Major efforts have been made in this field and a last algorithm is proposed" Multiple Part" for data analysis and extraction of new and more knowledge from them. In this paper, we propose a new version of Multiple Part algorithm which integrates a heuristic evaluation method and a feature elimination technique in order to extract multiple and equivalent solution for biomedical data. In order to prove the validity of our algorithm, we analyze an RNA-seq of cancer diseases data sets extracted from The Cancer Genome Atlas (TCGA). Furthermore, we validate our approach by comparing it with the existing methods. Experimental results show the efficacy of our proposed algorithm.**

***Index Terms*—Multiple solution, Multiple Part, Camur, Merit, heuristic method, IMPA.**

I. Introduction

The cancer mechanism becomes more worldwide major public health issue. Since cancer is one of the leading causes of mortality, many researches have been developed in order to understand its mechanisms and discover new knowledge to prevent and to treat this serious disease [1]. In recent years RNA-seq protocol counting the RNA fragments that are aligned on a reference genome. In this scenario, it is important to identify informative genes with high prognostic value to distinguish between healthy tissue and tumoral tissue types. In this work, we focus on the amelioration and the adoption of a new algorithm for classifying RNA-seq case-control samples, which is able to compute multiple human readable classification models. In the past, such problems have been solved by the use of supervised and unsupervised machine learning algorithms such as decision tree, rule-based, ensembles decision tree, neural networks and Support Vector Machines (SVM) [2] [3] [4] [5]. These techniques have been also used to improve diagnosis of diseases such as Alzheimer, Breast Cancer or Meningitis [6] [7] [1].The big limits with the application of these machine learning algorithms are related to the managing of the huge amount of data. In fact, for biological datasets, a high learning time is needed for data analysis and the extraction of new knowledge from them [8]. Also, all these classical algorithms compute just a single classification method that contains few of features. While our goal is the extraction and the discovery of the maximum knowledge from these RNA sequence datasets by computing many alternative and equivalent classification models.

Multiple and equivalent solutions extraction from biological datasets is a novel concept which has recently caught the attention of researchers. Obtaining a set of efficient solutions with a better compromise between the features and with a reduced running time is the goal of this study. More details about these works are presented in the rest of the paper. All these methods can on one side provide a relevant number of rules (solutions) with low performance. On the other side, the number of extracted rules at each iteration can be insufficient compared to the big RNA-sequence datasets used.

In this work, we propose a new algorithm to optimize Multiple Part algorithm for classifying RNA-seq case-control samples. This algorithm integrates a discretization method, a feature elimination technique and a heuristic evaluation method for each subset of selected features. The final aim of this work is to provide a more compact, human interpretable models that can aid biologists or doctors to make a decision about the classification of diseases. The rest of this article is organized as follows. In section 2 we present a literature review about methods to extract multiples and equivalent solutions. In section 3 we describe our proposed algorithm (IMPA). The experiments and their results are discussed in section 4. Finally, in section 5, We report the conclusion and we present future works.

II. Literature review

A number of previous studies have been focused on the extraction of multiple and equivalent solutions in biomedical data classification problems. One approach is presented in Fiscon et al, [9], where the authors proposed meta-heuristic approach based on an evolutionary algorithm to find a solution for identifying a large number of small species-specific genomic

DOI reference number: 10.18293/SEKE2021-137

subsequences. One other work proposed by Gholami et al., [10], this classification-based approach is based on recursive feature elimination RFE method. The limit of this algorithm is that at each iteration, only a single variable should be chosen to remove. This would be inefficient in many high dimensional applications such RNA-sequence datasets.

In recent years, several works pointed to extract multiple solutions interpretable by human using rule and tree-based classification algorithms. Valerio et al, [11], proposed a new algorithm Camur (Classifier with a Alternative and Multiple Rule-based model). This algorithm able to extract, multiple, alternative and equivalent rule-based models (Ripper). These rules represent the most relevant set of features related to the case and control samples. In 2016, Fiscon et al, [9], proposed a metaheuristic approach in order to find solutions for identifying a large number of small species-specific genomic subsequences. This approach aims to extract multiple solutions using rule and tree-based classification algorithms. In 2017, Fabrizio Celli, et al, [1] developed a new algorithm called Big Biomedical data classifier (BIGBIOCL). This algorithm able to classify a large DNA methylation dataset. BIGBIOCL is inspired by Camur algorithm in order to apply classification methods to big datasets. In 2019, Guannoni et al, [8] proposed a new method that extracts multiple and equivalent classification methods. This method Called" Multiple Part" algorithm that integrates rule-based classification method (Part) and a feature elimination technique in order to obtain more interpretable models in a reduced execution time. In the first, this method iteratively computes the rule-based classifier, then it computes the power set of the features present in the rules, iteratively eliminates these combinations from the data sets, and execute again the classification procedure until a stopping criterion is verified. Experimental results show that" Multiple part" is an important algorithm for extracting multiple, equivalent and alternative solution in a reduced execution time.

III. The proposed algorithm: Improved Multiple Part Algorithm (IMPA)

We propose an enhanced version of" Multiple Part algorithm" which specifies the quality of each combination of the features found in the rule using an heuristic evaluation method. We called our proposed method as IMPA (Improved Multiple Part Algorithm). IMPA is new algorithm inspired by" Multiple Part" in order to extract multiple and equivalent solutions with higher performance and in few time executions. It is a tool to obtain knowledge by extracting several alternative classification models for gene features in RNA-seq data. Through evaluation of the possible combination to delete, and through iterative deletion of selected features, extraction of equivalent classification models is possible using IMPA algorithm. The implementation of our new algorithm is essentially based on feature elimination method by evaluating each power set of features. One of the reasons is that the merit function for evaluation the set of features enables to evaluate the worth of a subset of features by considering the individual predictive ability of each feature as well as with the degree of redundancy between them.

A. Steps of the IMPA algorithm

IMPA implements the following steps:

1) Compute Rule-based method (PART): our algorithm executes at first a rule-based algorithm (Part) that extract a set of logic rules "if CONDITION then CLASS" rules which provide an immediate relationship between the class and one or more features (genes).
2) Computes the power set of the features present in the rule: IMPA calculates the power set of the features present in the rule after each iteration. Then, all the combination are stored in a memory list.
3) Discretize the data set: to computes the score of each combination, we need to discretize continuous features. A copy of the training data is first discretized then passed to compute the quality of each combination features. In this work we choose to use the discretization method of Fayyad and Irani [12] because it has been showed that the number of classification errors generated by this method is comparatively smaller than the number of errors generated by the other discretization algorithms.
4) Compute the quality of each combination features using merit function: we use a correlation based heuristic evaluation function for computing the score of each set of combination feature. This function called" Merit function". Merit function is a measure that calculates feature-class and feature-feature correlations using a measure called symmetrical uncertainty (SU) correlation. This function enables to evaluate the heuristic" merit" of feature subsets. It ranks the feature subsets according to a correlation based heuristic evaluation function [13]. The subset with the lowest merit is considered the first combination to be eliminated from the data set at time. let Pk is a subset of features (one combination), we define the Merit function associated with Pk as follows:
$$Merit(P^k) = \frac{j * \overline{r_{yx}}}{\sqrt{j + j * (j-1) * \overline{r_{xx}}}} \qquad (1)$$
where j=$|P^k|$ is the number of subset features, r_{yx} is the average of the correlations between the subset features and the class and r_{xx} is the average inter-correlation between subset features.
The numerator of Equation 1 can be considered to provide an indication of the predictive of the class a set of features are; The denominator represents how much redundancy there is among the features [13]. Merit function uses SU to measure correlation. SU [14] associated with two features x_1 and x_2 is defined by:
$$SU(x_1, x_2) = 2 * [\frac{GI}{H(x_1) + H(x_2)}] \qquad (2)$$
more details about the SU function is presented in [14]. The advantage of the Merit function is that is allows to

compare subsets of feature in different sizes. Thus, it allows to evaluate the contribution of a new feature.

5) Scores all possible features combination: after computing a score of each possible combination, we sort the list of combination in ascending order according to the score (The worst Merit to the best Merit).
6) Perform feature elimination method: eliminates all the possible combinations of features by starting with the worst Merit and run the analysis again at each time. The feature elimination is iterated in two execution-mode:
   - A loose feature elimination mode: in the first, a classification with the PART algorithm is performed. This mode takes the results from the first classification and build the combinations (power set) of the found features, whose combinations are iteratively eliminated according to the worst score from the data set. After each elimination of the feature combination, a classification step is built. The new extracted features that are present in the current classification model are added to the features list and are going to be processed in the next iterations. In loose mode, once a feature is removed it inserted again in the data set.
   - A strict feature elimination mode: in the first, a classification with the PART algorithm is performed. The features appear from the first model are extracted and then eliminated one by one according to the worst score. A classification is iterated after each elimination on the resulting data set. In the strict mode, once a feature is eliminated it is never inserted again in the data set.
7) Our proposed algorithm performs again the classification procedure until a stopping criterion is verified: the reliability (F-measure) < a given threshold, maximum number of iterations (Max-iter) is reached, or the list of features has been completely treated.

In the final, we obtain a several of relevant number of equivalent classification models e.g.," IF feature <1.50 then the sample is NORMAL" with higher performance. These rules composed of a list of relevant genes related to a particular class.

B. *Execution example of our IMPA algorithm*

Given a data set of RNA-seq data related to Breast cancer with two class tumoral and normal:

- IMPA extracts through the first execution a model composed of a set of rules, e.g.,$(ADHFE1 \geq 4.69) AND (ACSBG1 \geq 0.37) OR (HBBP1 \leq 0.04)$ then normal
- The rules contain a set of three features (genes) S1 ={ADHFE1, ACSBG1, HBBP1}.
- The power set is computed: P1 ={{ADHFE1}, {ACSBG1}, {HBBP1}, {ADHFE1, ACSBG1}, {ADHFE1,HBBP1}, {ACSBG1,HBBP1}, {ADHFE1, ACSBG1, HBBP1 }}.
- Discretize the dataset.
- Compute the quality of each combination features using merit function. P1= Merit {{ADHFE1}=0.7, P2= Merit{ACSBG1} =0.3, P3=Merit {HBBP1}=0.1, P4=Merit{ADHFE1, ACSBG1}=0.04, P5=Merit{ADHFE1,HBBP1}=1.18, P6=Merit {ACSBG1,HBBP1}=0.130, P7=Merit{ADHFE1, ACSBG1, HBBP1}=0.175
- Sort the list of combination in ascending order: {P4, P3, P6, p7, p5, P2, P1}.
- The first item of the power set is eliminated from the data set and the classification is performed, which provides a new set of features, S2 = {HBBP1, ADH4}.
- The first power set P1 is completely performed.
- After the treating of P1, the power set P2 from S2 is computed and the classification is performed.
- The algorithm continues again the classification procedure until one of the stopping criteria is verified.

IV. Experimental study and results

A. *Description of the dataset*

Our experimental analysis in focused on RNA-seq data related to Breast cancer disease (BRCA). These data are extracted from public available data of The Cancer Genome Atlas (TCGA) [15] [16]. The data set of BRCA composed of a matrix in comma separated value format, which is the input of our algorithm. The rows of the matrix correspond to a set 59 samples that represent the sequenced tissues of the patients. The columns correspond to 20532 features which represent the gene expression profile. The last column represents the class e.g., normal - tumoral. Each cell contains the gene expression measure Reads Per Kilobase per Million mapped reads (RPKM) value for each gene expression measure. [17].

TABLE I: Data matrix of breast cancer RNA-seq data

Sample ID	ANO8	Clorf27	TRPM6	....	Class
A8-A09D	2.64	5.42	0.38	....	Breast cancer
BH-A0DH	1.46	6.47	0.76	....	Normal
GM-A2D9	3.13	14.21	0.61	....	Breast cancer
....	....	....	....	....	....
GM-A2DB	3.86	5.15	0.59	....	Breast cancer

Concept and experimental study: We compare in the experimental study the obtained results of IMPA with results of CAMUR and" Multiple Part" [8]. Our comparison will be based on the number of extracted models, the performance of the models, the number of relevant features and the execution time. In fact, our goal is to validate a new supervised classification algorithm able to extract multiple models by building hundreds of classification iterations on a massive number of relevant features in few hours. We choose to variate the iteration numbers (Iter-nb) between 20 and 150. Also, we variate the minimum number of F-measure on 0.8 and 0.9. We use for each parameter the two-execution mode" strict" and" loose". For evaluating the classification models, we adopt the accuracy and the F -measure equations. –Our proposed algorithm is implemented in JAVA language programming. The experimentation has been executed on a laptop with an on 2.71 GHz Intel (R) Core (TM)i7 CPU and 32 GB of RAM. Table III reports the genes that are most represented in the rules. Table IV, Table V and Table VI represent the classification result, the number of extracted rule (Nb-rule), the number

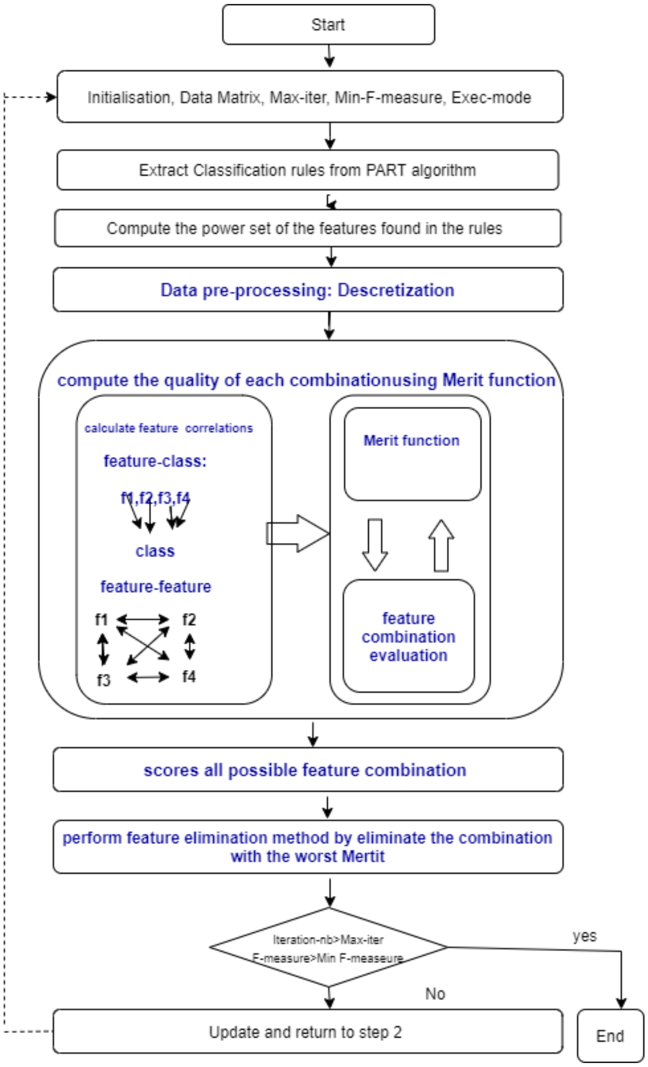
Start
Initialisation, Data Matrix, Max-iter, Min-F-measure, Exec-mode
Extract Classification rules from PART algorithm
Compute the power set of the features found in the rules
Data pre-processing: Descretization
compute the quality of each combinationusing Merit function
calculate feature correlations
feature-class:
f1,f2,f3,f4
class
feature-feature
f1
f2
f3
f4
Merit function
feature combination evaluation
scores all possible feature combination
perform feature elimination method by eliminate the combination with the worst Mertit
Iteration-nb>Max-iter
F-measure>Min F-measeure
yes
No
Update and return to step 2
End

Fig. 1: The process of IMPA algorithm.

TABLE II: Classification rules example extracted from Camur, multiple part and IMPA with a classification accuracy $\geq 90\%$.

Extracted rules of Camur	Extracted rules of Multiple Part	Extracted rules of IMPA
(ADH4 $\mid$ 127 $\geq$ 0.26) $\|$ (AHDC1 $\mid$ 27245 $\geq$ 11.02) $\Rightarrow normal$	(ACSM2B $\mid$ 348158 $\leq$ 0.02) AND (HBBP1 $\mid$ 3044 $\leq$ 0.04) $\Rightarrow BRC$	(ADHFE1$\mid$ 137872 $\geq$ 4.69) AND (ACSBG1 $\mid$23205 $\geq$ 0.37) OR (HBBP1 $\mid$ 3044 $\leq$ 0.04) $\Rightarrow normal$

of extracted features (Nb-f), the average accuracy (Aver-acc) of each rule and the running time of each classification.

V. Discussion

From the tables we can conclude some considerations regarding the link between the number of extracted rules, the number of features, the accuracy of the rules and the execution time. The running time of our IMPA algorithm is faster than Camur but not for the" multiple Part". Multiple Part remain the faster because it uses a faster method for extracted rules (Part) while our algorithm used the same method of extracted rules but it integrated more calculation instructions. Figure 2 show this difference of execution time.

TABLE III: The most represented genes in the rule using IMPA algorithm: extracted genes related to breast cancer.

Features	Occurence
RAG1AP1 55974	5
HOXA7 3204	5
CDC A855143	5
LOC 572558	3
RERGL 79785	3

From Table IV, when we variate the number of iterations between 20 and 80, all the iterations are executed for the three algorithms. Then, the number of extracted rules and the number of relevant genes for IMPA algorithm are higher compared to the other algorithms. Since all the iteration number are executed for all the algorithms, the execution time of IMPA is less than Camur but not less to Multiple Part. In addition, by varying the iteration number between 100 and 150, not all the iteration are executed for Camur , so a small number of accurate rules are extracted in these analyses. By comparing Multiple Part and IMPA, all the iteration are executed but the number of accurate rules (accuracy 0.8) of IMPA is larger compared to Multiple Part. Therefore, the number of relevant features is larger than Multiple Part. Hence, our IMPA algorithm can produce more higher equivalent classification models with higher accuracy than the other algorithms. Figure 4 shows this difference in term of extracted rules. We can explain this difference by the heuristic evaluation method that we have integrated for our algorithm before handle the feature elimination method. This method enables to evaluate the" Merit" of each combination of features to be eliminated.

On the other hand, as shown in Figure 3, The accuracy of extracted rules of IMPA in all cases is in the range [0.99, 1]. This mean that it extract always compact rules with the higher performance compared to the other algorithms.

As shown in Table III, the most represented genes extracted from the rules are RAG1AP1 with id 55974, HOXA7 with id 3204 and CDCA with id A855143. These genes are the most involved in the breast cancer classification models. Many studies have shown that HOXA7 plays a critical role in regulating the proliferation of estrogen receptor -positive cancer cells [18]. A recent study shows that RAG1AP1 is the new biomarker candidate of breast cancer development [19]. Another study shows that CDCA plays a crucial role for the prevention of this disease [20]. We can conclude that such information in the extracted rules IMPA can be considered as an important result to help biologists and doctors in analyzing the genetics of breast cancer disease.

Using loose feature elimination mode (Table V), all the algorithms completed all the iterations but they extract only a few numbers of extracted rules. The cause can be that after the first execution, the extracted rules do not exceed the f-measure value (0.8). In Table V, since almost all the algorithms provide

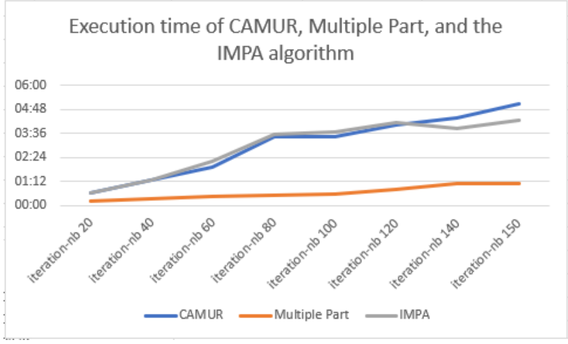

Fig. 2: Execution time of Camur, Multiple Part and IMPA.

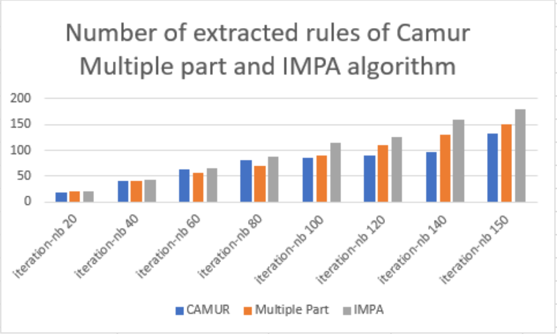

Fig. 4: Extracted rule number of Camur, Multiple Part, IMPA

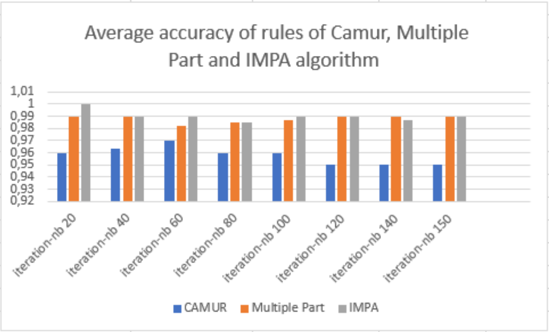

Fig. 3: Average Accuracy of extracted rules of Camur, Multiple part and IMPA

a few numbers of rules, our IMPA algorithm provides more rule number compared to the other algorithm and there are not classification errors compared to CAMUR. In Table VI the number of iterations is not treated for CAMUR because the stopping criteria is reached (f-measure smaller than 0.9). For multiple Part and IMPA, all the iteration are executed but they give all a few numbers of rules. The cause can be that the extracted rules does not contain the features and therefore the power set list to be removed is empty for each iteration. Thereby, IMPA algorithm provides more several efficient classification rules with high performance and remains faster then the other algorithms.

From this detailed analysis we can conclude that our IMPA algorithm is an elegant algorithm that is able to extract more multiple classification models with high accuracy for the RNA-sequence classification problem. Therefore, it enables to identify most of the features related to the investigated class. Our algorithm operating efficiently because the integration of the heuristic method to evaluate the feature subset which is based it can help to provide more accuracy human interpretable models. Moreover, our algorithm can be efficient and can provide thousands of equivalent solutions in one single run.

VI. Conclusion

In this work, we presented a new algorithm (IMPA) enables to extract multiple and equivalent models for RNA-sequence classification problem. Our proposed algorithm adopted a feature elimination technique and integrated an heuristic evaluation method for each subset of selected features in order to provide more accuracy rules for each classification model. IMPA is applied on a set of RNA-seq data focusing on Breast cancer from TCGA. After the experimental study, we prove that our proposed algorithm is a reliable technique for extract more compact rules with more relevant features than multiple Part and CAMUR. It can also ignore redundant and duplicate rules to be executed when ordered and evaluates the power set of features to be eliminated.

In a future work, we plan to more ameliorate the execution time of our algorithm to be applied on big data set. As another future work we can extend the analyses to another biological data set, e.g., RNA-sequece data of COVID-19, DNA-methylation values and DNA-Barcoding in order to confirm the validity of our approach. Also, we are investigating the possibility to validate the extracted genes by domain experts with deep analysis.

References

[1] F. Celli, F. Cumbo, and E. Weitschek, "Classification of large dna methylation datasets for identifying cancer drivers," *Big Data Research*, 2018.

[2] E. Weitschek, G. Fiscon, and G. Felici, "Supervised dna barcodes species classification: analysis, comparisons and results," *BioData mining*, vol. 7, no. 1, p. 4, 2014.

[3] M. Elloumi and A. Y. Zomaya, *Biological Knowledge Discovery Handbook: Preprocessing, Mining and Postprocessing of Biological Data*. John Wiley & Sons, 2013, vol. 23.

[4] N. Guannoni, R. Sassi, W. Bedhiafi, and M. Elloumi, "A comparison between classification algorithms for postmenopausal osteoporosis prediction in tunisian population," in *International Conference on Information Technology in Bio-and Medical Informatics*. Springer, 2016, pp. 234–248.

[5] F. Previtali, P. Bertolazzi, G. Felici, and E. Weitschek, "A novel method and software for automatically classifying alzheimer's disease patients by magnetic resonance imaging analysis," *Computer methods and programs in biomedicine*, vol. 143, pp. 89–95, 2017.

[6] G. D'Angelo, R. Pilla, C. Tascini, and S. Rampone, "A proposal for distinguishing between bacterial and viral meningitis using genetic programming and decision trees," *Soft Computing*, vol. 23, no. 22, pp. 11 775–11 791, 2019.

[7] A. A. Mohamed, W. A. Berg, H. Peng, Y. Luo, R. C. Jankowitz, and S. Wu, "A deep learning method for classifying mammographic breast density categories," *Medical physics*, vol. 45, no. 1, pp. 314–321, 2018.

[8] N. Guannoni, F. Mhamdi, E. Weitschek, and M. Elloumi, "Novel algorithm to extract multiple solutions for rna sequence classification problem," in *2019 International Conference on High Performance Computing & Simulation (HPCS)*. IEEE, 2019, pp. 856–863.

TABLE IV: Results of classification analysis with Camur, Multiple Part and IMPA using Strict execution mode (F-measure=0.8).

Parameters	**Camur**					**Multiple Part**					**proposed method**				
Strict mode,F-measure=0.8	# iter	Nb-rules	Aver acc	Nb-f	Exec /min	# iter	Nb-rules	Aver acc	Nb-f	Exec /min	# iter	Nb-rules	Aver acc	Nb-f	Exec /min
Iter-nb=20	20	18	0.96	22	00:38	20	20	0.99	20	00:14	20	22	1	25	00:38
Iter-nb=40	40	40	0.963	47	01:18	40	40	0.99	40	00:20	40	43	0.99	48	01:16
Iter-nb=60	60	64	0.97	82	01:54	60	57	0.982	63	00:26	60	66	0.99	71	02:13
Iter-nb=80	80	80	0.96	96	03:25	80	69	0.985	82	00:32	80	87	0.985	96	03:33
Iter-nb=100	85	86	0.96	101	03:28	100	90	0.987	103	00:35	100	115	0.99	130	03:40
Iter-nb=120	88	91	0.95	111	04:01	120	110	0.99	123	00:47	120	125	0.99	134	04:10
Iter-nb=140	100	97	0.95	113	04:24	140	130	0.99	143	01:06	140	160	0.987	173	03:50
Iter-nb=150	140	132	0.95	167	5:05	140	150	0.99	153	01:52	150	180	0.99	183	04:16

TABLE V: Results of classification analysis with Camur, Multiple Part and IMPA using loose execution mode (F-measure=0.8).

Parameters	**Camur**					**Multiple Part**					**proposed method**				
loose mode, F-measure =0.8	# iter	Nb-rules	Aver acc	Nb-f	Exec /min	# iter	Nb-rules	Aver acc	Nb-f	Exec /min	# iter	Nb-rules	Aver acc	Nb-f	Exec /min
Iter-nb=20	20	3	0.972	3	00:21	20	2	1	2	00:09	6	5	0.99	4	00:19
Iter-nb=40	40	2	1	2	00:17	40	2	1	2	00:09	5	4	1	3	00:18
Iter-nb=60	60	2	0.958	2	00:16	60	2	1	2	00:09	5	5	0.99	3	00:14
Iter-nb=80	80	2	0.958	2	00:15	80	2	1	2	00:09	6	5	1	4	00:22
Iter-nb=100	100	3	0.972	3	00:18	100	2	1	2	00:09	8	7	0.99	6	00:30
Iter-nb=120	120	4	1	4	00:16	120	2	1	2	00:09	6	5	1	4	00:22
Iter-nb=140	140	4	0.958	4	00:15	140	2	1	2	00:09	6	5	1	4	00:18
Iter-nb=150	150	3	0.972	3	00:17	160	2	1	2	00:09	4	3	1	2	00:10

TABLE VI: Results of classification analysis with Camur, Multiple Part and IMPA using Strict execution mode (F-measure=0.9).

Parameters	**Camur**					**Multiple Part**					**proposed method**				
Strict mode, F-measure =0.9	# iter	Nb-rules	Aver acc	Nb-f	Exec /min	# iter	Nb-rules	Aver acc	Nb-f	Exec /min	# iter	Nb-rules	Aver acc	Nb-f	Exec /min
Iter-nb=20	20	22	0.97	22	00:47	20	20	1	20	00:14	20	22	1	25	00.61
Iter-nb=40	24	28	0.982	32	00:46	40	47	0.985	40	00:19	40	43	0.99	48	01.16
Iter-nb=60	20	21	0.98	21	00:47	60	57	0.981	63	00:27	60	64	0.981	70	01.20
Iter-nb=80	46	40	0.97	60	01:29	80	69	0.985	82	00:33	80	80	0.99	88	01:34
Iter-nb=100	52	53	0.98	70	02:16	100	91	0.981	103	00:45	100	110	0.99	125	02:05
Iter-nb=120	63	77	0.97	92	02:37	120	110	0.996	123	00:46	120	123	0.996	133	03:21
Iter-nb=140	49	46	0.96	53	02:10	140	132	0.99	143	00:51	140	150	0.996	169	02:40
Iter-nb=150	68	67	0.968	78	03:10	150	144	0.992	153	01:17	150	177	0.993	179	02:43

[9] G. Fiscon, E. Weitschek, E. Cella, A. L. Presti, M. Giovanetti, M. Babakir-Mina, M. Ciotti, M. Ciccozzi, A. Pierangeli, P. Bertolazzi *et al.*, “Missel: a method to identify a large number of small species-specific genomic subsequences and its application to viruses classification,” *BioData mining*, vol. 9, no. 1, p. 38, 2016.

[10] B. Gholami, I. Norton, A. R. Tannenbaum, and N. Y. Agar, “Recursive feature elimination for brain tumor classification using desorption electrospray ionization mass spectrometry imaging,” in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*. IEEE, 2012, pp. 5258–5261.

[11] V. Cestarelli, G. Fiscon, G. Felici, P. Bertolazzi, and E. Weitschek, “Camur: Knowledge extraction from rna-seq cancer data through equivalent classification rules,” *Bioinformatics*, vol. 32, no. 5, pp. 697–704, 2015.

[12] U. Fayyad and K. Irani, “Multi-interval discretization of continuous-valued attributes for classification learning,” 1993.

[13] M. A. Hall, “Correlation-based feature selection for machine learning,” 1999.

[14] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, “Numerical recipes in c,” 1988.

[15] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, C. G. A. R. Network *et al.*, “The cancer genome atlas pan-cancer analysis project,” *Nature genetics*, vol. 45, no. 10, 2013.

[16] F. Cumbo and E. Weitschek, “An in-memory cognitive-based hyperdimensional approach to accurately classify dna-methylation data of cancer,” in *International Conference on Database and Expert Systems Applications*. Springer, 2020, pp. 3–10.

[17] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, “Mapping and quantifying mammalian transcriptomes by rna-seq,” *Nature methods*, vol. 5, no. 7, p. 621, 2008.

[18] Y. Zhang, J.-C. Cheng, H.-F. Huang, and P. C. Leung, “Homeobox a7 stimulates breast cancer cell proliferation by up-regulating estrogen receptor-alpha,” *Biochemical and biophysical research communications*, vol. 440, no. 4, pp. 652–657, 2013.

[19] M. P. Świtnicki, M. Juul, T. Madsen, K. D. Sørensen, and J. S. Pedersen, “Pincage: probabilistic integration of cancer genomics data for perturbed gene identification and sample classification,” *Bioinformatics*, vol. 32, no. 9, pp. 1353–1365, 2016.

[20] N. N. Phan, C.-Y. Wang, K.-L. Li, C.-F. Chen, C.-C. Chiao, H.-G. Yu, P.-L. Huang, and Y.-C. Lin, “Distinct expression of cdca3, cdca5, and cdca8 leads to shorter relapse free survival in breast cancer patient,” *Oncotarget*, vol. 9, no. 6, p. 6977, 2018.