Future Computing and Informatics Journal

Volume 6
Issue 2 (2021) *Issue 2*

Article 1

2021

A Statistical-Mining Techniques' Collaboration for Minimizing Dimensionality in Ovarian Cancer Data

Mohamed Attia
Faculty of Computers and Information Technology, Future University in Egypt,
Mohamed.abdelgawad@fue.edu.eg

Maha Farghaly
Faculty of Computers and Information Technology, Future University in Egypt, Maha.farghaly@fue.edu.eg

Mohamed Hamada
Faculty of Computers and Information Technology, Future University in Egypt,
Mohamed.hamada@fue.edu.eg

Amira M. Idrees AMI
Faculty of Computers and Information Technology, Future University in Egypt,
amira.mohamed@fue.edu.eg

Future Computing and Informatics Journal

Volume 6
Issue 2 (2021) *Issue 2*

Article 1

2021

A Statistical-Mining Techniques' Collaboration for Minimizing Dimensionality in Ovarian Cancer Data

Mohamed Attia
Faculty of Computers and Information Technology, Future University in Egypt,
Mohamed.abdelgawad@fue.edu.eg

Maha Farghaly
Faculty of Computers and Information Technology, Future University in Egypt, Maha.farghaly@fue.edu.eg

Mohamed Hamada
Faculty of Computers and Information Technology, Future University in Egypt,
Mohamed.hamada@fue.edu.eg

Amira M. Idrees AMI
Faculty of Computers and Information Technology, Future University in Egypt,
amira.mohamed@fue.edu.eg

Recommended Citation

Attia, Mohamed; Farghaly, Maha; Hamada, Mohamed; and Idrees, Amira M. AMI (2021) "A Statistical-Mining Techniques' Collaboration for Minimizing Dimensionality in Ovarian Cancer Data," *Future Computing and Informatics Journal*: Vol. 6 : Iss. 2 , Article 1.
DOI: http://doi.org/10.54623/fue.fcij.6.2.1
Available at: https://digitalcommons.aaru.edu.jo/fcij/vol6/iss2/1

Future Computing and Informatics Journal, Vol. 6 [2021], Iss. 2, Art. 1
Future Computing and Informatics Journal
Homepage: https://digitalcommons.aaru.edu.jo/fcij/
doi: http://org/10.45623/fue.fcij.6.2.1

A Statistical-Mining Techniques' Collaboration for Minimizing Dimensionality in Ovarian Cancer Data

Mohamed Attia [1, a], Maha Farghaly[1, b], Mohamed Hamada[1, c], Amira M. Idrees[1, d]

[1]Faculty of Computers and Information Technology, Future University in Egypt
[a] Mohamed.abdelgwad@fue.edu.eg, [b] Maha.farghaly@fue.edu.eg,
[c] Mohamed.hamada@fue.edu.eg, [d] amira.mohamed@fue.edu.eg

ABSTRACT
A feature is a single measurable criterion to an observation of a process. While knowledge discovery techniques successfully contribute to many fields, however, the extensive required data processing could hinder the performance of these techniques. One of the main issues in processing data is the dimensionality of the data. Therefore, focusing on reducing the data dimensionality through eliminating the insignificant attributes could be considered one of the successful steps for raising the applied techniques' performance. On the other hand, focusing on the applied field, ovarian cancer patients continuously suffer from the extensive analysis requirements for detecting the disease as well as monitoring the treatment progress. Therefore, identifying the most significant required analysis could be a positive step to reduce the emotional and financial suffering. This research aims to reduce the data dimensionality of the ovarian cancer disease and highlight the most significant analysis using the collaboration of clustering techniques and statistical techniques. The research succeeded to identify twelve significant analysis out of forty-four with a total of fourteen significant attributes for ovarian cancer data.

Keywords
Data Dimensionality; Saaty Method; Clustering; Classification; Ovarian Cancer; Weighting Techniques

https://digitalcommons.aaru.edu.jo/fcij/vol6/iss2/1
DOI: http://doi.org/10.54623/fue.fcij.6.2.1

1. INTRODUCTION

A feature is a single measurable criterion to an observation of a process. Artificial intelligence (AI) makes use of the feature criterion in any machine learning algorithm especially in clustering as well as classification problems [1]. In gene data analysis or medical related problems, feature selection techniques are typically used to find the most informative genes and how differentially expressed genes are discovered [2]. However, the main concern of feature selection is to focus on a subset of variables; representing the features; which can effectively manipulate the input data while minimizing the redundancy and removing the irrelevant attributes as well as noisy data if exists [3]. Based on the expansion of the applications of machine learning or pattern recognition, the domain of features has been significantly increased from tens to hundreds of attributes or features applied in those applications [4]. As different research targeted the health sector [5], it is a fact that several approaches are developed to point the problem of minimizing irrelevant and redundant dimensions. For that, feature selection or attribute minimization keeps track of all the given information (input data) and tries to analyze, understand data, and extract the important attributes targeting to minimizing the effect of dimensionality to improve prediction effectiveness [6].

The feature or gene data in the medical field maintains a very high probability of containing numerous numbers of attributes, in addition to that, most of these attributes probably participate with other attributes in a correlation manner (e.g., when two features are perfectly correlated, only one feature is sufficient to describe the data). So, the dependent attributes provide no further details about the labels, also can be represented as a noise over the prediction. All the available information can be extracted and maintained through minimum number of features. These features uniquely correspond to the whole data which contains maximum gain of information about the labels. Hence by excluding these dependent attributes, the data size which effect the prediction process will be reduced, thus the effectiveness of the approach increased [7].

In other cases, not only noisy attributes; which have no relation to the labels; slow the process, but also may have the ability to bias data curve to certain point. This situation reduces the algorithm performance as there is no sufficient information about the process being studied. In addition, many researchers studied the main approaches of feature selection as listed below [1]:

1. Filtering Approach

This approach makes use of the attribute ranking mechanisms as the core principle criteria for feature selection through ordering. These methods are maintained for their flexibility and simplicity which have been applicable for practical problems. A suitable ordering criterion used to score the attributes based on a threshold value which will be an evidence to exclude all attributes below target value. Each feature is evaluated separately through its statistical criteria's, there is no specific model used. Thus, it is independent of the classifier. The most typically used filter methods are:

- Mutual Information
- Information Gain (IG)
- Minimum Redundancy Maximum
- Correlation based Feature Selection (CFS)
- Fast correlation-based filter (FCBF)
- Scoring algorithm
- Random Forest Ranking

Published by Arab Journals Platform, 2021

2. Wrapper Approach

Wrapper approach considers the predictor and performance as a black box and evaluation function respectively to measure the attributes' subset. Since evaluating the whole set(s) is represented as a NP-hard problem, so an optimal subset can be performed based on search algorithms which find a subset heuristically. A lot of search algorithms can be used to find a subset of variables to maximize the evaluation function gain for a given machine learning problem. In classification, it involves using learning mechanisms to extract the core subset. This approach merges the classifier with the search space which leads to more accuracy achieved. First, it begins with a sample of the solution, and this sample; which is the features' subset; is measured. Based on the fitness function, the wrapper approach risks are computation cost and overfitting, the following techniques are mainly representing the wrapper approach goal,

- Sequential selection algorithms
- Genetic Algorithm (GA)
- Artificial Bee Colony (ABC)
- Ant Colony Optimization (ACO)
- Particle Swarm Optimization
- Black Hole Algorithm (BHA)
- Harmony Search algorithm (HSA)

3. Hybrid Approach

The hybrid or ensembled approach is based on making use of filter and wrapper approaches. The combination of both approaches achieves high efficiency through two stages; minimizing the feature space dimension based on ranking or filtering then obtaining the wrapper method to pick the optimal feature subset. However, the performance accuracy still not guaranteed due to the differences of the mechanisms of both approaches separately. In addition, the ensemble approach proposes that the merge process of multi-experts is better than single output expert, a single wrapper approach accuracy is not guaranteed for different datasets [1] [6].

Therefore, AI data driven solutions for health care sector, especially cancer, are considered a challenging study and open research area for most of researches [8] [3]. Many of these researches try to study the impact of artificial intelligence on health care [9] [10], especially on cancer detection or prediction [11]. The emergence of feature (gene) selection has encouraged researchers to analyze data with high dimensions for reducing dimensionality. The cancer became one the most critical diseases and the correlated data hold numerous information with dependent and independent attributes. consequently, intelligent techniques are required to extract the most important features with extreme gain to the application's performance. This paper aims to investigate the most influencing attributes for ovarian cancer disease data. The remaining of the paper will introduce the related work, the research significance, the proposed approach, and the experimental study results.

2. RELATED WORK

In this literature the research goes through identifying different studies of features selection techniques to detect and identify different diseases. The authors of [13] stated the importance of detecting gynaecological cancer stage to decide the treatment plan. The stage indicates the degree of spread and exacerbation of the cancer. They intended to enhance the prediction of gynaecological cancers stages, specifically cervical and ovarian cancers. The work proposed a framework named Revised and Improved Feature Subset through Fused Feature Selection (RIFSt_2FS). The framework is applied on a dataset of cancer patients' data who are diagnosed with cervical or ovarian

https://digitalcommons.aaru.edu.jo/fcij/vol6/iss2/1
DOI: http://doi.org/10.54623/fue.fcij.6.2.1

cancer between 2000 and 2017. The dataset was conditionally selected from The SEER (Surveillance, Epidemiology, and End Results Program) database. The framework is a crossbreeding of filter feature selection method and wrapper feature selection method, which are Relief Algorithm and Genetic Algorithm, respectively. The work applied the Random Forest classifier for cancer stage prediction using the selected features achieving accuracy of 97%.

Another research in [14] believed that the early discovery of ovarian cancer increases survival chance. Their literature focused on simplifying the prediction of ovarian cancer. The proposed framework is applied on a dataset of 349 Chinese patients with 49 variables which is collected between 2011 and 2018 from the Third Affiliated Hospital of Soochow University. The dataset was partitioned into 68% for training and 32% for testing. The Literature achieved the state of the art by applying the Minimum Redundancy - Maximum Relevance (MRMR) feature selection method for 48 times to each attribute in the dataset. MRMR tends to recognize the features with the highest power of discrimination between the targeted classes with considering the existence of correlation between features. MRMR succeeded to identify ten significant features which are then used to train the classification model. The literature employed CART algorithm to build a decision tree for prediction which resulted in finding the most two significant features. The authors compared the results accuracy replacing MRMR with reliefF feature selection. The comparison results highlighted that MRMR selected more significant features with higher F1 score of classification than reliefF feature selection. The authors also approved that using CART decision tree for classification reached the best AUC-ROC score when compared to using the risk of ovarian malignancy algorithm (ROMA) or logistic regression for classification.

Moreover, the literature in [15] considered not only the high dimensionality of data, but also the heterogeneous nature of multi-omics data. A TCGA (The Cancer Genome Atlas) dataset of multi-omics ovarian cancer are gene level copy number alteration (CNA), DNA methylation, and gene expression (GE) RNA-Seq, with total of 13877 omics features the dataset was exported and pre-processed from UCSC Xena cancer genomic browser and used on the study. The study proposed a double consequent staged framework to combine and homogenize multi-omics data. The first stage contains single view filters that can be any conventional feature selection methods. Each filter is associated with a switch to be set if the input data is a single view data. The second stage contains only one filter that could be a multi-view feature selection method in case the method will be applied directly to the input, otherwise, the input specific view data will be concatenated as an input to stage2 filter. The work also proposed an extension to Minimum Redundancy - Maximum Relevance (MRMR) feature selection method for multi-omics data. The extension was named multi view Minimum Redundancy - Maximum Relevance (mv-MRMR). The mv-MRMR idea was to capture the selected features for a specific view that represent the maximal relevant features among all views. The work results demonstrated that predicting ovarian cancer using multi-view data is much powerful and accurate than using specific-view data. The results also approved that mv-MRMR outperforms single-view models.

Published by Arab Journals Platform, 2021

Additionally, the authors of [16] indicated a limitation of the support vector machine recursive feature elimination procedure (SVM-RFE), that is, it tries recursively to find the best combinations for binary classification. The research proposed sigFeature, a feature selection algorithm, which was based on support vector machine and t- statistic. The research argued that the proposed algorithm overcomes the indicated limitation. The literature testing was performed on six microarray data sets of six different types of cancer available on Gene Expression Omnibus, one of them was a dataset of 195 samples of ovarian cancer patients. The literature compared sigFeature to a three of state-of-the-art algorithms, which are Support Vector Machine Recursive Feature Elimination (SVM-RFE), Support Vector Machine T-Test Recursive Feature Elimination algorithm (SVM-T-RFE) and Support Vector Machine Bayesian T-Test Recursive Feature Elimination algorithm (SVM-BT-RFE). The work approved that sigFeature not only selects the most accurate significant features for prediction, but also recognizes the biological signature of the dataset among SVM-RFE, SVM-T-RFE, and SVM-BT-RFE. The signature of the selected features was validated using gene set enrichment analysis (GSEA) using the Molecular Signatures Database (MSigDB).

More research in [17] focused on the blood analysis molecularly or atomically which confirmed that it is a promising test to diagnose ovarian cancer that could overcome the limitations of ultra-sound imaging and cancer antigen CA-125 tests. A dataset of 176 patients' blood plasma labelled as either normal, ovarian cyst or ovarian cancer, partitioned into one third for validation and two thirds for training. Principal component analysis (PCA) scores were used for selecting the validation samples. The training partition was used to extract the significant features in order to predict ovarian cancer. For the feature selection phase, SelectKBest algorithm and chi-squared test were employed. For prediction, a regression model based on back-propagation neural network was trained with adopting five-fold cross validation. The model validation is performed on random selected instances. The prediction achieved sensitivity of 71.4% and specificity of 86.5%. The limitation of blood plasma analysis is the time consumption of collecting the dataset.

3. RESEARCH SIGNIFICANCE

The exhaustive requirements for medical examinations of the ovarian cancer detection as well as the follow up consequently lead to suffering from both physical and financial bottlenecks. These bottlenecks have been arisen due to the high cost of many of these examinations as well as the need for continuous medical visits. Following this lead has highlighted the current research direction. This research targets the minimization of the required indications which highlight the alarm for detecting ovarian cancer as well as the follow up of the disease progress. The main research direction is to explore the weight of each examination contribution in the detection and follow up process.

The examinations are considered as a set of attributes which provide a full view for the patient case. Following the research approach leads to minimizing the required examinations by setting the most significant examinations' set which consequently leads to minimizing the examinations cost for the patient as he will seek for less examinations;' set. On the other hand, following the research

https://digitalcommons.aaru.edu.jo/fcij/vol6/iss2/1
DOI: http://doi.org/10.54623/fue.fcij.6.2.1

approach also leads to preserving the patient's physical conditions to its most possible status by avoiding the examinations which proved to be less contributing in the detection as well as the follow up process.

The following points are listed as a summary for the main research significance:

1. The research highlights the fact that each weighting technique has its own perspective, therefore, the collaboration between different techniques provides a wider perspective and consequently, a more accurate weighting result.
2. The contribution of the weighting techniques is based on evaluating each technique individually by following the adapted Saaty method in [22]. the authors claim that this evaluation phase provides more accurate attributes' weighting.
3. As it is confirmed by many researches that the nature of the data has its significant effect on the selected techniques, therefore, exploring the consistent weighting techniques' set will be performed based on the nature of the ovarian cancer dataset.
4. On the other hand, the consistent weighting techniques are applied targeting to explore the most significant attributes' set for ovarian cancer data.
5. The final attributes' exploration phase includes applying clustering techniques for highlighting the correlations between the contributing attributes and ensures high accuracy in the exploration phase.
6. The classification technique is then applied as a validation phase for exploring the significant attributes and ensure the correct classification after truncating the attributes' set which represents the final
7. significant examinations for ovarian cancer detection and follow up.

4- THE PROPOSED APPROACH FOR EXPLORING THE MOST SIGNIFICANT EXAMINATIONS FOR OVARIAN CANCER

Measuring the consistency level for each attribute is introduced following the adapted Saaty method in [22] as a primary path. However, further adaptation in [23] has been evaluated to provide high classification accuracy. The proposed adapted method in [23] introduced extending the parameters' set in Saaty method to become ten parameters while it was only three parameters in [22]. While the main scope of [22] [23] was applying adapted Saaty method, in the current research, an additional step is proposed as a preliminary step targeting for higher accurate results. This research introduces applying a clustering step to ensure the ability to apply the adapted Saaty method on the applicable number of attributes. As the maximum applicable attributes for original Saaty was ten attributes and although this criterion was not highlighted in [23], however, this research follows applying the adapted Saaty method with a maximum of ten attributes. Although the research focuses on ovarian dataset, however, figure 1 illustrates the proposed general steps for minimizing the data dimensionality, while the following formulas identify the main contributing resources of the proposed method followed by the algorithmic steps of the proposed method.

Published by Arab Journals Platform, 2021

Input:

Weighting techniques' set Weight_Tech= {$wtech_1$, $wtech_2$,….$wtech_n$ | n ∈ ℕ }

Dataset attributes: Dset_Att = {A_1, A_2,….A_m| m ∈ ℕ }

Processing:

Clust_Att = {C_i, ….C_j}

{ <A_e,…A_q>,…., <A_a,…A_s> | A_e, A_q, A_a, A_s ∈ Dset_Att }

| Clust_Att| = | Dset_Att|

S_i = Clust_Att (C_i) = {A_e,…A_q}

Output:

The consistent attributes' set Consistent_ATT = {A_f,….A_k | A_f , A_k ∈ Dset_Att }

Consistent_ATT ⊆ Dset_Att

Algorithm:

For each cluster C_i

{ For each attribute A_j where A_j ∈ S_i

{ for each weighting technique $wtech_n$ ∈ Weight_Tech

{

Apply $wtech_n$ on S_i with label A_j

Apply adapted Saaty method for A_j

Identify consistency (A_j, $wtech_n$)

}

Identify consistency_Status (A_j)

}

}

Determine consistent attributes' set members (Consistent_ATT) where

Consistent_ATT = { A_j} where consistency_Status (A_j) = consistent

https://digitalcommons.aaru.edu.jo/fcij/vol6/iss2/1
DOI: http://doi.org/10.54623/fue.fcij.6.2.1

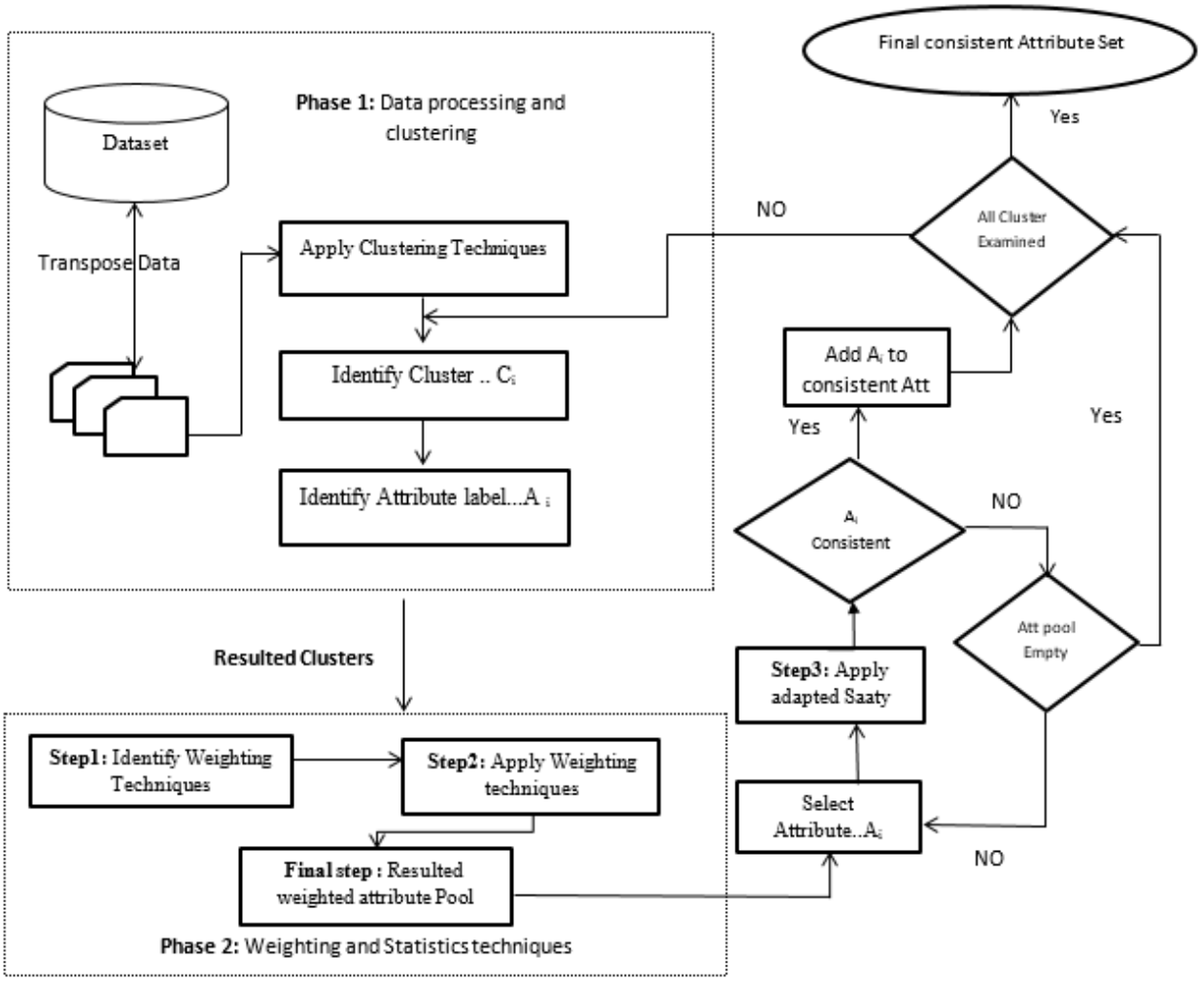

Figure 1: The proposed general steps for data dimensionality

5- EXPERIMENTAL CASE STUDY (OVARIAN CANCER DATA)

A dataset of ovarian cancer patients is the focus of the research's case study. The dataset was available by [24], which targeted the prediction of ovarian cancer. The dataset has a total of 350 patients' records while 261 of them are ovarian cancer patients' records and 139 are benign ovarian tumors' patients. The target of using this dataset is to identify the most significant features that could discriminate between the two types of patients. The dataset records are described by a total of 47 features which represent the patient medical description including demographics, blood routine test, general chemistry, and tumor markers. Table 1 illustrates a description of the describing features while table 2 illustrates a sample of the patients' records where type 1 is an ovarian cancer patient record while type 2 is a benign ovarian tumors' patient record.

Published by Arab Journals Platform, 2021

Table 1: Ovarian Cancer Dataset contributing attributes [24]**.**

Abbreviation	Biomarker Name	Sample type	Instrument	Method	Lower limit	Upper Limit	Unit
MPV	Mean platelet volume	full blood	Sysmex XE-2100	Calculation method	7.4	12.5	fL
BASO#	Basophil Cell Count	full blood	Sysmex XE-2100	FCM with semiconductor laser	0	0.06	10^9/L
PHOS	phosphorus	serum	Beckman Coulter AU5800	Phosphomolyase method	0.7	1.62	mmol/l
GLU.	glucose	serum	Beckman Coulter AU5800	Glucose oxidase method	3.9	6.1	mmol/l
CA72-4	Carbohydrate antigen 72-4	serum	Roche Cobas 8000	ECLIA	0	7	U/ml
K	kalium	serum	Beckman Coulter AU5800	ion selective electrode method	3.5	5.3	mmol/l
AST	Aspartate aminotransferase	serum	Beckman Coulter AU5800	MDH method	6	40	u/l
BASO%	Basophil Cell ratio	full blood	Sysmex XE-2100	Calculation method	0	1	%
Mg	magnesium	serum	Beckman Coulter AU5800	XB-I method	0.73	1.3	mmol/l
CL	chlorine	serum	Beckman Coulter AU5800	ion selective electrode method	99	110	mmol/l
CEA	Carcinoembryonic antigen	serum	Roche Cobas 8000	ECLIA	0	5	ng/ml
EO#	eosinophil count	full blood	Sysmex XE-2100	FCM with semiconductor laser	0.02	0.52	10^9/L
CA19-9	Carbohydrate antigen 19-9	serum	Roche Cobas 8000	ECLIA	0	37	U/ml
ALB	albumin	serum	Beckman Coulter AU5800	bromcresol green method	35	55	g/l
IBIL	Indirect bilirubin	serum	Beckman Coulter AU5800	Calculation	2	15	umol/L
GGT	Gama glutamyltransferasey	serum	Beckman Coulter AU5800	Rate method	3	73	u/l
MCH	Mean corpuscular hemoglubin	full blood	Sysmex XE-2100	Calculation method	27	34	Pg
GLO	globulin	serum	Beckman Coulter AU5800	Calculation	20	40	g/l
ALT	Alanine aminotransferase	serum	Beckman Coulter AU5800	Lactate dehydrogenase method	1	45	u/l
DBIL	direct bilirubin	serum	Beckman Coulter AU5800	Vanadate oxidation	1.5	7	umol/l
RDW	red blood cell distribution width	full blood	Sysmex XE-2100	Calculation method	10.6	15.5	%
PDW	Platelet distribution width	full blood	Sysmex XE-2100	Calculation method	15.5	18.1	%
CREA	creatinine	serum	Beckman Coulter AU5800	Jaffe method	44	144	umol/l
AFP	alpha-fetoprotein	serum	Roche Cobas 8000	ECLIA	0	7	ng/ml
HGB	hemoglobin	full blood	Sysmex XE-2100	SLS-Hemoglobin method	110	150	g/l
Na	Natrium	serum	Beckman Coulter AU5800	ion selective electrode method	137	147	mmol/l
HE4	human epididymis protein 4	serum	Roche Cobas 8000	ECLIA	0	140	pmol/L
LYM#	lymphocyte count	full blood	Sysmex XE-2100	FCM with semiconductor laser	1.1	3.2	10^9/L
CA125	Carbohydrate antigen 125	serum	Roche Cobas 8000	ECLIA	0	35	U/ml
BUN	blood urea nitrogen	serum	Beckman Coulter AU5800	Urease-glutamic acid dehydrogenase	1.7	8.3	mmol/l
LYM%	lymphocyte ratio	full blood	Sysmex XE-2100	Calculation method	20	50	%
Ca	calcium	serum	Beckman Coulter AU5800	Arsenazo III method	1.12	1.32	mmol/l
AG	Anion gap	serum	Beckman Coulter AU5800	Calculation	8	30	mmol/l
MONO#	mononuclear cell count	full blood	Sysmex XE-2100	FCM with semiconductor laser	0.1	0.6	10^9/L
PLT	platelet count	full blood	Sysmex XE-2100	Hydrodynamic Focusing DC method	125	350	10^9/L
NEU	neutrophil ratio	full blood	Sysmex XE-2100	Calculation method	40	75	%
EO%	eosinophil ratio	full blood	Sysmex XE-2100	Calculation method	0.02	0.52	10^9/L
TP	Total protein	serum	Beckman Coulter AU5800	Biuret method	60	82	g/l
UA	urie acid	serum	Beckman Coulter AU5800	Urease method	90	450	μmol/l
RBC	Red blood cell count	full blood	Sysmex XE-2100	Hydrodynamic Focusing DC method	3.5	5.5	10^12/L
PCT	thrombocytocrit	full blood	Sysmex XE-2100	Calculation method	0.114	0.282	L/L
CO2CP	carban dioxide-combining Power	serum	Beckman Coulter AU5800	Enzymatic method	18	30	mmol/l
TBIL	total bilirubin	serum	Beckman Coulter AU5800	Vanadate oxidation	4	19	μmol/l
HCT	hematocrit	full blood	Sysmex XE-2100	Hydrodynamic Focusing DC method	0.35	0.45	L/L
MONO%	monocyte ratio	full blood	Sysmex XE-2100	Calculation method	3	10	%
MCV	mean corpuscular volume	full blood	Sysmex XE-2100	Calculation method	82	100	fL
ALP	Alkaline phosphatase	serum	Beckman Coulter AU5800	NPP substrate-AMP buffer method	25	130	u/l

https://digitalcommons.aaru.edu.jo/fcij/vol6/iss2/1
DOI: http://doi.org/10.54623/fue.fcij.6.2.1

Table 2: A sample of Ovarian Cancer patients' Dataset [24] **(transposed)**

TYPE	0	0	0	0	1	1	1	1
MPV	11.7	10	11.4	7.38	9.4	6.76	11.4	10.6
BASO#	0.01	0.02	0.03	0.05	0.01	0.06	0	0.02
PHOS	1.46	1.09	0.97	1.25	1.2	1.14	1.01	1.42
GLU.	4.67	10.5	4.64	4.76	5.8	4.71	4.46	4.46
K	5.36	4.38	4.3	4.7	4.33	4.59	4.06	4.05
AST	24	13	18	17	24	10	12	19
BASO%	0.3	0.3	0.6	0.74	0.2	0.72	0.6	0.7
Mg	0.78	0.82	1	1.11	1.21	0.98	1.05	1.11
Menopause	0	1	0	1	0	0	0	0
CL	107.4	100.1	102.6	103.2	97.4	99.3	96.5	100.2
CEA	1.4	2.46	0.77	0.82	1.9	1.29	0.9	1.38
EO#	0.04	0.04	0.03	0	0.06	0	0	0.04
CA19-9	36.48	19.98	12.18	18.41	11.66	12.44	9.65	14.22
ALB	45.4	39.9	45.4	39.2	48.3	41.2	44	34.8
IBIL	3.5	4.2	10.1	8	4.1	3.2	5.7	6.3
GGT	16	13	10	17	20	11	11	18
MCH	33.7	26.2	28.4	30.6	29.4	30.1	31.2	28.6
GLO	28.5	32.1	32.5	26.9	28.7	27.8	35.5	37.8
ALT	11	9	9	16	28	13	11	18
DBIL	2	2.6	4.7	2.9	2.5	2.5	2.9	3.2
Age	47	61	39	45	30	26	28	32
RDW	13.7	12.7	12	14.6	12.3	14	13	12.5
PDW	13.4	11.2	15.2	17.4	10.4	16.5	16.3	11.5
CREA	103	45	48	65.7	74.5	60.3	49.2	58
AFP	3.58	34.24	1.5	2.75	0.61	4.16	0.61	0.77
HGB	89	128	131	123	151	128	129	129
Na	141.3	142	138.9	139.1	140.4	136.3	138.9	139.1
HE4	208.5527	934.1	47.56	853.5	47.61	38.79	42.2	40.94
LYM#	0.65	1.27	1.1	1.73	2.13	1.33	1.8	1.33
CA125	15.36	2444	56.08	2555	69.13	35.98	17.8	13.99
BUN	5.35	3.21	3.8	5.27	5.31	3.18	3.75	4.3
LYM%	16.8	17.2	23.7	27.2	36	17.2	39.6	43.54
Ca	2.48	2.62	2.57	2.35	2.43	2.56	2.47	2.45
AG	19.36	23.98	18.4	16.6	22.63	17.09	23.76	20.75
MONO#	0.22	0.41	0.25	0.42	0.37	0.3	0.3	0.37
PLT	74	304	112	339	227	219	213	173
NEU	76.2	76.5	69.7	65.5	56.5	78.2	53.3	42.44
EO%	1	0.5	0.6	0.07	1	0	0.4	1.3
TP	73.9	72	77.9	66.1	77	69	79.5	72.6
UA	396.4	119.2	209.2	215.6	195.9	157.7	258	167.3
RBC	2.64	4.89	4.62	4.01	5.13	4.27	4.14	4.51
PCT	0.09	0.3	0.13	0.25	0.21	0.148	0.243	0.18
CO2CP	19.9	22.3	22.2	24	24.7	24.5	22.7	22.2
TBIL	5.5	6.8	14.8	10.9	6.6	5.7	8.6	9.5
HCT	0.273	0.417	0.391	0.372	0.457	0.388	0.374	0.386
MONO%	5.7	5.5	5.4	6.55	6.3	3.83	6.1	12.14
MCV	103.4	85.3	84.6	92.6	89.1	91	90.4	85.6
ALP	56	95	77	26	76	33	67	51

Clustering the attributes' set has been applied using k-means algorithm with initiating the number of clusters to be five clusters [25]. This research have followed the proposed approach in [23] in order to identify the most suitable number of clusters. In order to perform the required attributes' clustering, the dataset has been transposed and the k-means algorithm has been applied, table 3

illustrates the five clusters' attributes' members.

Table 3: Attributes' Clusters Members

	No. of members	Attributes
cluster 0	20	BASO#, BASO%, BUN, Ca, CEA, DBIL, EO#, EO%, GLU., HCT, K, LYM#, Menopause, Mg, MONO#, MONO%, MPV, PCT, RBC, PHOS
cluster 1	14	AFP, AG, ALT, AST, CA72-4, CO2CP, GGT, GLO, IBIL, LYM%, MCH, PDW, RDW, TBIL
cluster 2	11	Age, ALB, ALP, CA19-9, CL, CREA, HGB, MCV, Na, TP, NEU
cluster 3	2	CA125, HE4
cluster 4	2	PLT, UA

Following the proposed approach in [23], the research tackled the argument the consistency of the weighting measures follows different criteria including the applied methods' approach as well as the nature of data. The research in [23] examined a set of weighting methods against the Gastrology data, the set included "Principle Component Analysis (PCA), Information Gain, Information Gain Ratio, Support Vector Machine (SVM), Gini Index, Chi Square, Deviation, and Correlation". The result of the examination highlighted five to six out of the eight measures to be consistent with the Gastrology dataset, they are "Information Gain, Information Gain Ratio, Support Vector Machine, Deviation, Chi-Square, and Correlation". Referring to the attributes' types, the authors reached a conclusion that the Gastrology dataset has the same medical data nature as the ovarian cancer dataset. Therefore, the current research adopts the same weighting measures subset as consistent for ovarian cancer patients' data. Consequently, the current research adopts the same approach in applying the six consistent weighting measures. Following the proposed approach, the attributes' consistency has been measured using the weighting measures that belong to the consistent measures' set. As previously mentioned, the set was described by 49 attributes after excluding the patients' ID attribute. According to the experiment results, table 4 illustrates the weight for each attribute using the consistent weighting measures' set members.

Demonstrating the original attributes' weight has been performed to highlight the effect of applying the proposed approach targeting to explore the consistent attributes as the most significant attributes. The next stage is determining the attributes' consistency, the proposed method for consistency exploration has been applied on each cluster members individually. As previously highlighted that the adapted Saaty method follows the same nature of the original Saaty which states that the method is applicable up to 10 attributes [22]. The following presents the results for applying the stage steps over "Menopause" attribute as an example for applying the stage steps.

To explore the consistency for each attribute, the attribute under examination is set to be the label attribute and the weighting for the other cluster members is performed for the six clusters. Following the example on focus, "Menopause" attribute has been identified as the label attribute, then the weighting measures have been applied on the remaining cluster members. Table 4 presents the adapted weight for the contributing attributes which are the members of cluster 0 (19 attributes).

https://digitalcommons.aaru.edu.jo/fcij/vol6/iss2/1
DOI: http://doi.org/10.54623/fue.fcij.6.2.1

Table 4: Weight for cluster 0 attributes' members with considering the "Menopause" attribute as the label attribute.

	IGR	deviation	chi squared	correlation	IG	SVM	PCA
Mg	0.08	0.01	0.02	0.02	0	0.07	-0.01
BASO#	0.03	0	0.03	0.09	0.01	0.06	0.01
DBIL	0.08	0.01	0.04	0.07	0.01	0.01	-0.01
CEA	0.08	0.02	0.03	0.04	0.06	0.01	0.01
MONO%	0.16	0.01	0.02	0.03	0.02	0.01	0.01
MONO#	0.08	0.01	0.03	0.12	0.01	0.01	0.01
EO#	0.07	0.01	0.02	0.1	0.02	0.08	-0.01
HCT	0.16	0.01	0.04	0.06	0.01	0.02	0.01
K	0.18	0.01	0.05	0	0.02	0.05	0.01
EO%	0.1	0.01	0.02	0.11	0.03	0.06	-0.01
RBC	0.18	0.01	0.04	0.07	0.03	0	0.01
PHOS	0.16	0.01	0.01	0.07	0.01	0.09	-0.01
PCT	0.09	0.01	0.03	0.18	0.02	0.02	0.01
BASO%	0.08	0.01	0.04	0.17	0.04	0.02	-0.01
MPV	0.18	0.01	0.04	0.05	0.01	0.14	-0.01
Ca	0.1	0.01	0.06	0.22	0.05	0.07	-0.01
LYM#	0.16	0.01	0.05	0.23	0.05	0.07	-0.01
BUN	0.19	0.01	0.12	0.34	0.11	0.23	0.01
GLU.	0.24	0.01	0.15	0.4	0.13	0.16	0.01

It is illustrated in table 3 that Menopause attribute is a member in cluster 0 which included 20 attributes. As the research follows the adapted Saaty method, which is limited to 10 attributes, therefore, the highest weighted attributes in the same cluster have been selected for exploring the consistency status. Therefore, the consistency of Menopause attributes has been examined with the contribution of a subset of cluster 0 which has the highest weight with respect to the weighting measure on focus as the following clarifying the applied process. Table 5 presents these contributing attributes for each weighting Technique.

Table 5: Contributing attributes in each weighting measure with "Menopause" attribute as the label attribute.

Weighting Measure	Contributing Attribute
Information Gain Ratio	Ca, MONO%, HCT, PHOS, LYM#, K, RBC, MPV, BUN, GLU.
Information Gain	K, PCT, EO%, RBC, BASO%, Ca, LYM#, CEA, BUN, GLU.
Chi-Square	DBIL, HCT, RBC, BASO%, MPV, K, LYM#, Ca, BUN, GLU.
Correlation	BASO#, EO#, EO%, MONO#, BASO%, PCT, Ca, LYM#, BUN, GLU.
Support Vector Machine	BASO#, EO%, Mg, Ca, LYM#, EO#, PHOS, MPV, GLU., BUN
Deviation	RBC, PHOS, PCT, BASO%, MPV, Ca, LYM#, BUN, GLU., CEA

Following the presented attributes' set that is illustrated in table 6, the adapted Saaty method has been applied as proposed in [23], ten measure have contributed to exploring the "Menopause" attribute consistency status, they are the minimum, maximum, mean, median, range, upper inter-quartile range, lower inter-quartile range, mean of upper inter-quartile range and the mean value, and mean of lower inter quartile range and the mean value of λ. Finally, the same consistency threshold is below or equal 0.3 which has been identified as the maximum acceptable percentage for consistency [22].

According to these weights, the proposed adapted Saaty method has been applied and the attribute's consistency has been measured. The following groups of illustrations presents different

Published by Arab Journals Platform, 2021

situations. Table 6 illustrates the values for the adapted Saaty method parameters of "Menopause" attribute and the contribution of the highest weighted ten attributes whish are members in the same cluster with respect to "information gain ratio" weighting Technique. According to the results in table 6, then table 7 illustrates the λ Measures for which concluded the Menopause" attribute consistency with respect to information gain ration weighting Technique while figure 2 presents a graphical distribution of the λ' measures. The presented λ' measures values in table 7 highlighted the fact that the consistency of the Menopause" attribute is neutral as illustrated in table 8 that five of the λ' measures are consistent while five of them are not.

Table 6: adapted Saaty method parameters of "Menopause" attribute WRT Information Gain Ratio Weighting Technique

IGR	1	2	3	4	5	6	7	8	9	10	Product	Eigen Vector
1	1.00	0.38	0.38	0.38	0.19	0.13	0.30	0.30	0.17	0.38	**0.31**	**0.03**
2	2.07	1.00	1.00	1.00	0.50	0.33	0.80	0.89	0.45	1.00	**0.80**	**0.07**
3	2.07	1.00	1.00	1.00	0.50	0.33	0.80	0.89	0.45	1.00	**0.80**	**0.07**
4	2.07	1.00	1.00	1.00	0.50	0.33	0.80	0.89	0.45	1.00	**0.80**	**0.07**
5	5.33	2.00	2.00	2.00	1.00	0.67	1.60	6.70	0.89	2.00	**1.88**	**0.16**
6	8.00	3.00	3.00	3.00	1.50	1.00	2.40	2.67	1.34	3.00	**2.47**	**0.21**
7	3.33	1.25	1.25	1.25	0.63	0.42	1.00	9.00	0.56	1.25	**1.27**	**0.11**
8	3.00	1.13	1.13	1.13	0.56	0.38	0.90	1.00	0.50	1.13	**0.93**	**0.08**
9	6.00	2.25	2.25	2.25	1.13	0.75	1.80	2.00	1.00	2.25	**1.85**	**0.16**
10	2.07	1.00	1.00	1.00	0.50	0.33	0.80	0.89	0.45	1.00	**0.80**	**0.07**
											11.90	

Table 7: λ Measures for Menopause" attribute WRT Information Gain Ratio Weighting Technique

	L	C.I	C.R	Consitency
Lmax	14.0955823	0.4550647	0.78459432	**Not Const.**
Lmean	10.7877072	0.08752302	0.15090176	**Const.**
Lmedian	10.3113509	0.03459455	0.05964578	**Const.**
Lmin	10.2755728	0.03061919	0.05279171	**Const.**
range	6.82000958	-0.3533323	-0.6091936	**Not Const.**
StDev	1.21556483			
interQuartile range	14.4344017	0.4927113	0.84950224	**Not Const.**
	7.14101267	-0.3176653	-0.5476987	**Not Const.**
mean(rang,UpperQ)	10.6272056	0.06968951	0.12015434	**Const.**
mean(rang,LowerQ)	6.98051113	-0.3354988	-0.5784461	**Not Const.**
mean(range, mean)	8.80385838	-0.1329046	-0.2291459	**Const.**

https://digitalcommons.aaru.edu.jo/fcij/vol6/iss2/1
DOI: http://doi.org/10.54623/fue.fcij.6.2.1

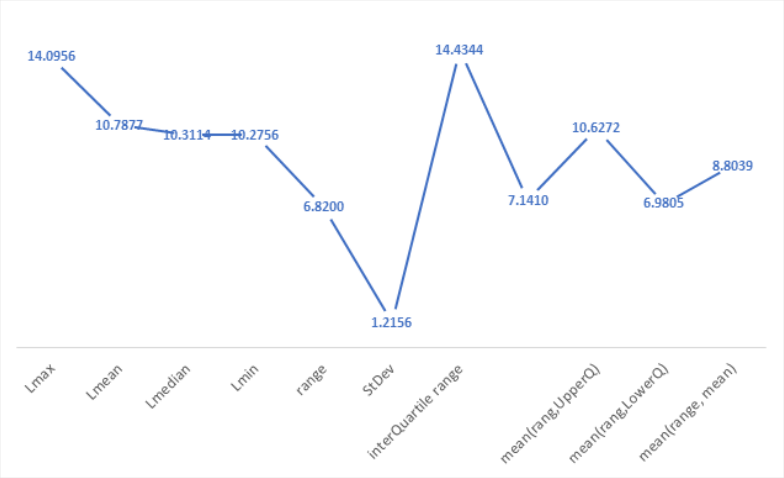

Figure 2: Distribution of λ Measures for "Menopause" attribute WRT Information Gain Ratio Weighting Technique

Table 8: Count and Percentage of Consistency status in λ' measures

	Count	Percentage
Consistent λ' measures	5	50%
Non-Consistent λ' measures	5	50%

The similar steps have been applied for each weighting technique, for more clarification, tables 9, 10, and 11 illustrate the same sequence of steps for the "Menopause" attribute is consistency status with respect to the deviation weighting technique with graphical illustration of λ Measures distribution in figure 3. The presented results highlight the fact that the "Menopause" attribute is consistent as illustrated in table 9 that six of the λ' measures are consistent while four of them are not.

Table 9: adapted Saaty method parameters of "Menopause" attribute WRT Deviation Weighting Technique

Deviation	1	2	3	4	5	6	7	8	9	10	Product	Eigen Vector
1	1.00	1.00	1.00	1.00	1.00	1.00	0.10	0.11	0.06	0.13	**0.39**	**0.02**
2	1.00	1.00	1.00	1.00	1.00	1.00	0.10	0.11	0.06	0.13	**0.39**	**0.02**
3	1.00	1.00	1.00	1.00	1.00	1.00	0.10	0.11	0.06	0.13	**0.39**	**0.02**
4	1.00	1.00	1.00	1.00	1.00	1.00	0.10	0.11	0.06	0.13	**0.39**	**0.02**
5	1.00	1.00	1.00	1.00	1.00	1.00	0.10	0.11	0.06	0.13	**0.39**	**0.02**
6	1.00	1.00	1.00	1.00	1.00	1.00	0.10	0.11	0.06	0.13	**0.39**	**0.02**
7	10.00	10.00	10.00	10.00	10.00	10.00	1.00	1.11	0.56	1.25	**3.88**	**0.20**
8	9.00	9.00	9.00	9.00	9.00	9.00	0.90	1.00	0.50	1.13	**3.49**	**0.18**
9	18.00	18.00	18.00	18.00	18.00	18.00	1.80	2.00	1.00	2.25	**6.98**	**0.35**
10	8.00	8.00	8.00	8.00	8.00	8.00	0.80	0.89	0.44	1.00	**3.10**	**0.16**
											19.78	

Table 10: λ Measures for Menopause" attribute WRT Deviation Weighting Technique

	L	C.I	C.R	Consitency
Lmax	10	0	0	**Const.**
Lmean	10	0	0	**Const.**
Lmedian	10	0	0	**Const.**
Lmin	10	0	0	**Const.**
range	3	-0.7777778	-1.3409962	**Not Const.**
StDev	0			
interQuartile range	10	0	0	**Const.**
	10	0	0	**Const.**
mean(rang,UpperQ)	6.5	-0.3888889	-0.6704981	**Not Const.**
mean(rang,LowerQ)	6.5	-0.3888889	-0.6704981	**Not Const.**
mean(range, mean)	6.5	-0.3888889	-0.6704981	**Not Const.**

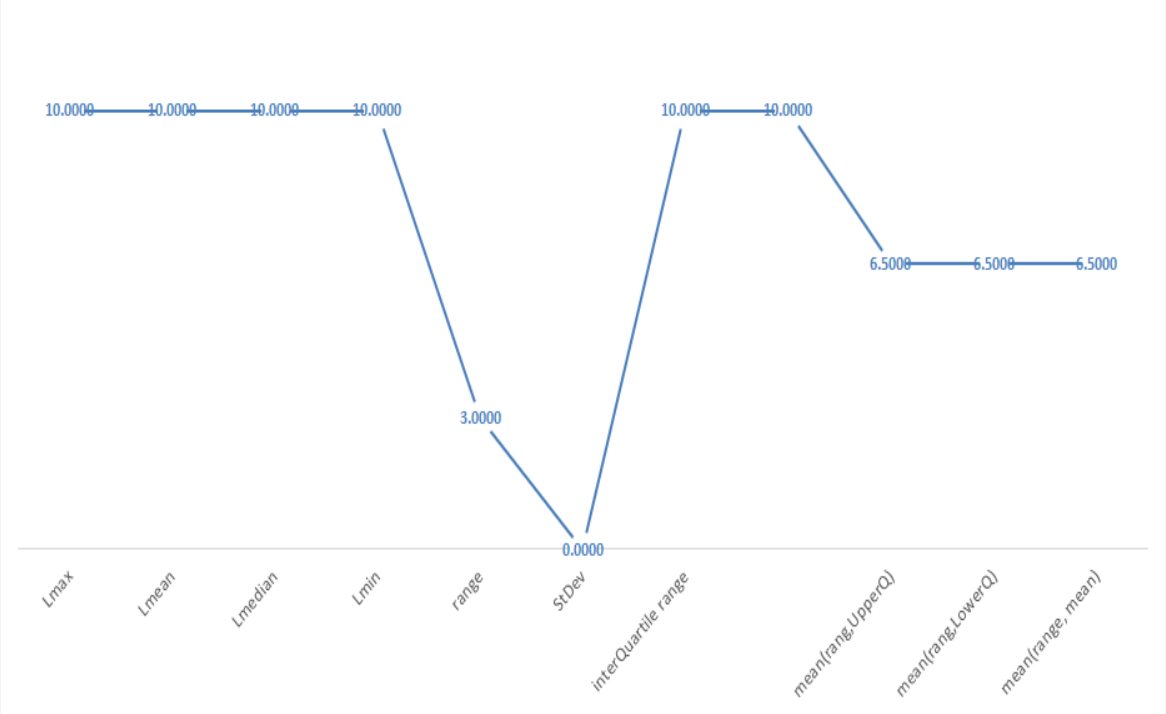

Figure 3: Distribution of λ Measures for "Menopause" attribute WRT Deviation Weighting Technique

Table 11: Count and Percentage of Consistency status in λ' measures

	Count	Percentage
Consistent λ' measures	6	60%
Non-Consistent λ' measures	4	40%

Moreover, as a final example, tables 13, 14, and 15 illustrate the same sequence of steps for the "Menopause" attribute is consistency status with respect to the deviation weighting technique with graphical illustration of λ Measures distribution in figure 4. The presented results highlight the fact that the "Menopause" attribute is not consistent as illustrated in table 15 that four of the λ' measures are consistent while six of them are not.

https://digitalcommons.aaru.edu.jo/fcij/vol6/iss2/1
DOI: http://doi.org/10.54623/fue.fcij.6.2.1

Table 12: adapted Saaty method parameters of "Menopause" attribute WRT Support Vector Machine Weighting Technique

SVM	1	2	3	4	5	6	7	8	9	10	Product	Eigen Vector
1	1.00	3.00	6.00	6.00	3.00	0.38	0.86	3.00	0.86	0.75	**1.70**	**0.14**
2	0.33	1.00	2.00	2.00	1.00	0.13	0.29	1.00	0.11	0.25	**0.52**	**0.04**
3	0.17	0.50	1.00	1.00	0.50	0.06	0.14	0.50	0.06	0.13	**0.26**	**0.02**
4	0.17	0.50	1.00	1.00	0.50	0.06	0.14	0.50	0.06	0.13	**0.26**	**0.02**
5	0.33	1.00	2.00	2.00	1.00	0.13	0.29	1.00	0.11	0.25	**0.52**	**0.04**
6	2.67	8.00	16.00	16.00	8.00	1.00	2.29	8.00	0.89	2.00	**4.12**	**0.34**
7	1.17	3.50	7.00	7.00	3.50	0.44	1.00	3.50	0.39	0.88	**1.80**	**0.15**
8	0.33	1.00	2.00	2.00	1.00	0.13	0.29	1.00	0.11	0.25	**0.52**	**0.04**
9	1.17	3.50	7.00	7.00	3.50	0.44	1.00	3.50	1.00	0.88	**1.98**	**0.17**
10	0.17	0.50	1.00	1.00	0.50	0.06	0.14	0.50	0.06	1.00	**0.32**	**0.03**
											11.99	

Table 13: λ Measures for "Menopause" attribute WRT Support Vector Machine Weighting Technique

	L	C.I	C.R	Consitency
Lmax	8.68035137	-0.1466276	-0.2528063	Const.
Lmean	8.57064173	-0.1588176	-0.2738234	Const.
Lmedian	8.68035137	-0.1466276	-0.2528063	Const.
Lmin	7.9256362	-0.2304849	-0.3973877	Not Const.
range	3.75471517	-0.6939205	-1.1964147	Not Const.
StDev	0.23754424			
interQuartile range	9.28327445	-0.0796362	-0.1373037	Const.
	7.85800901	-0.237999	-0.4103431	Not Const.
mean(rang,UpperQ)	6.51899481	-0.3867784	-0.6668592	Not Const.
mean(rang,LowerQ)	5.80636209	-0.4659598	-0.8033789	Not Const.
mean(range, mean)	6.16267845	-0.4263691	-0.7351191	Not Const.

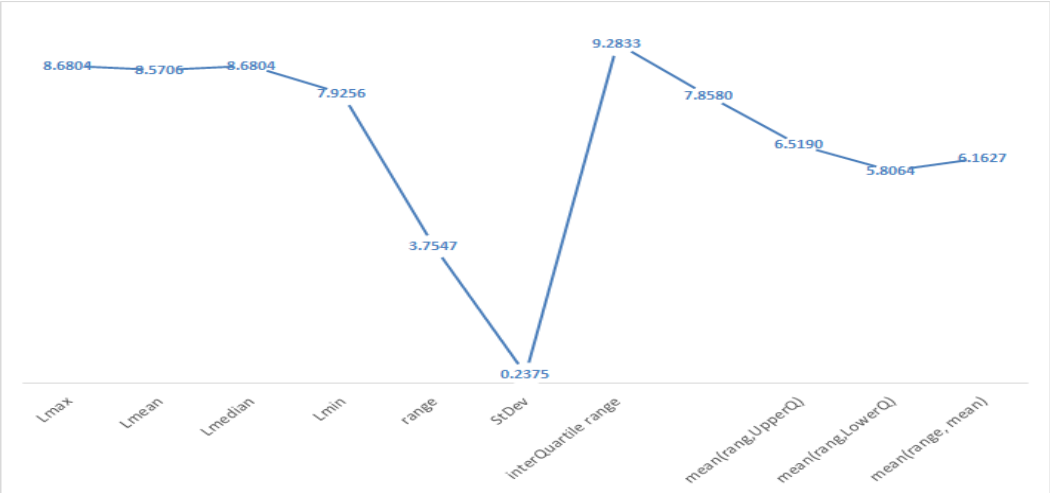

Figure 4: Distribution of λ Measures for "Menopause" attribute WRT Support Vector Machine Weighting Technique

Table 14: Count and Percentage of Consistency status in λ' measures

	Count	Percentage
Consistent λ' measures	4	40%
Non-Consistent λ' measures	6	60%

Published by Arab Journals Platform, 2021

Finally, table 16 presents the final decision considering the consistency status of the "Menopause" attribute. As shown in table 16, the "Menopause" attribute is consistent for three weighting technique, neutral for two weighting technique and not consistent for only one technique. Therefore, the final decision is considering the "Menopause" attribute to be consistent. This decision reveals that the "Menopause" attribute will be considered in the final attributes' set for the ovarian cancer patients' data.

Table 15: Weighting Techniques' Consistency Status for "Menopause" attribute

Information Gain (IG)	Information Gain Ratio (IGR)	Support Vector Machine (SVM)	Chi Square	Deviation (DEV)	Correlation (C)	Final Decision
Consistent	Neutral	Not Consistent	Neutral	Consistent	Consistent	Consistent

The same process is then performed for the remaining attributes targeting to reach the final attributes' set. Each of the attributes is examined following the same process with respect to its own cluster members as attributes for each cluster are considered in the exploration phase with eliminating other attributes without examination, the attribute has the sign C if it is consistent and the sign NC if it is not consistent. According to the consistency status for the attributes' set members, the final attributes set is determined to include all consistent attributes as illustrated in table 18 to be 26 attributes out of 49 attributes.

previously discussed for the "Menopause" attribute. Table 17 illustrates the final decision for all attributes. It is worth highlighting that only the highest weighted ten

Table 16: Consistency status for all attributes

Cluster 0				Cluster 1		Cluster 2	
Mg	NC	RBC	NC	TBIL	NC	HGB	NC
BASO#	NC	PHOS	NC	ALT	NC	TP	NC
DBIL	NC	PCT	C	MCH	NC	CL	C
CEA	NC	BASO%	NC	IBIL	C	CA19-9	C
MONO%	NC	MPV	C	GLO	C	PDW	NC
MONO#	NC	Ca	C	AFP	C	CREA	NC
EO#	C	LYM#	C	AG	C	MCV	C
HCT	NC	BUN	C	GGT	C	ALB	NC
K	C	GLU.	NC	CO2CP	NC	ALP	NC
EO%	C	Menopause	C	LYM%	C	NEU	C
Cluster 3		**Cluster 4**		RDW	NC	Na	C
HE4	C	PLT	C	AST	C	Age	C
CA125	C	UA	NC	CA72-4	C		

Table 17: Final Attributes' set members

EO#	K	MPV	LYM#	Menopause	AST	AG	GLO	IBIL	CL	NEU	Age	HE4
PCT	EO%	Ca	BUN	LYM%	CA72-4	GGT	AFP	MCV	CA19-9	Na	PLT	CA125

As a validation step, the ID3 classification technique [26] and KNN are applied on the dataset after eliminating the non-consistent attributes to ensure that the elimination step provides accurate classification with no effect on the accuracy as illustrated in table 19. The selection of ID3 and KNN for the classification task follows the research in [27] which compared the performance of a set of techniques for medical data. Although the research in [27] applied the techniques on Tinnitus patients' dataset and reached a conclusion that KNN and neural networks have the highest accuracy, however, the current research applied KNN and ID3 to confirm this conclusion.

https://digitalcommons.aaru.edu.jo/fcij/vol6/iss2/1
DOI: http://doi.org/10.54623/fue.fcij.6.2.1

Table 18: Performance measures for KNN classification Technique

Evaluation measure	Value
accuracy	81.58%
Classification error	18.42%
kappa	0.587
recall	88.20%,
precision	77.17%,
F-Score	82.31%

6. Conclusion

This research aimed to apply a set of steps targeting to highlight the most significant analysis for ovarian cancer disease. The research approach was to identify the highest weighted attributes in the ovarian cancer patients' dataset. The dataset was a benchmark which originally included ninety four attributes which describe the set of analysis for the patients. A total of three hundred and fifty records were included in the dataset, it was divided into two hundred and sixty one ovarian cancer patients while the remaining records were benign ovarian tumors. The research also highlighted the applicability of the successful collaboration between statistical and mining techniques for the required attributes' exploration task. The proposed method succeeded in minimizing the set of analysis into twenty four with a percentage of 42% attributes' minimization. Although the current research succeeded to reach its aim, however, more research directions could be suggested. Applying the proposed method on different datasets with different nature is one of the directions. Moreover, inspecting different clustering techniques could be further enhancement to confirm the heist performance.

References

[1] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers and Electrical Engineering,* vol. 40, no. 1, pp. 16-28, 2014.

[2] A. E. Khedr, A. Darwish, A. Z. Ghalwash and M. A. Osman, "Computer-Aided Early Detection Diagnosis System of Breast Cancer with Fuzzy Clustering Means Approach," *International Journal of Cancer Research,* vol. 48, no. 2, pp. 1257-1252, 2014.

[3] A. E. Khedr, A. Khalil and M. A. Osman, "Enhanced Liver Tumor Diagnosis Using Data Mining and Computed Tomography (CT)," *The International Conference on Computing Technology and Information Management (ICCTIM),* 2014.

[4] A. Khedr, S. Kholeif and F. Saad, "An Integrated Business Intelligence Framework for Healthcare Analytics," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 7, no. 5, pp. 263-270, 2017.

[5] M. Hazman and A. M. Idrees, "A Healthy Nutrition Expert System for Children," in *The 5th IEEE International Conference on E-Health and Bioengineering - EHB 2015,* 2015.

Published by Arab Journals Platform, 2021

[6] P. Agrawal, H. . F. Abutarboush, T. Ganesh and A. W. Mohamed, "Metaheuristic Algorithms on Feature Selection: A Survey of One Decade of Research (2009-2019)," in *IEEE Access 9*, 2021.

[7] B. Xue, M. Zhang, W. N. Browne and X. Yao, "A Survey on Evolutionary Computation Approachesto Feature Selection," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 4, pp. 606 - 626, 2016.

[8] M. A. Osman, A. Darwish, A. E. Khedr, A. Z. Ghalwash and A. E. Hassanien, "Enhanced breast cancer diagnosis system using fuzzy clustering means approach in digital mammography," in *Handbook of Research on Machine Learning Innovations and Trends*, IGI Global, 2017, pp. 925-941.

[9] M. M. Reda, Y. Helmy, A. E. Khedr and A. Abdo, "Intelligent Decision Framework to Explore and Control Infection of Hepatitis C Virus," in *International Conference on Advanced Machine Learning Technologies and Applications*, 2018.

[10] T. Sultan, A. Khedr, M. Nasr and R. Abdou, "A Proposed Integrated Approach for BI and GIS in Health Sector to Support Decision Makers," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 1, pp. 170-176, 2013.

[11] S. A. Taie and A. M. Idrees, "A Prototype for Breast Cancer Detection and Development Probability Expert System – Towards a Supportive Tool," in *The 5th IEEE International Conference on E-Health and Bioengineering - EHB 2015*, 2015.

[12] N. Almugren and H. Alshamlan, "A Survey on Hybrid Feature Selection Methodsin Microarray Gene Expression Data forCancer Classification, *IEEE Access*, " 2019.

[13] B Nithya and V Ilango, "Optimized Machine Learning based Classifications of Staging in Gynecological Cancers using Feature Subset through Fused Feature Selection Process," *International Journal of Advanced Computer Science and Applications (IJACSA)*, 2020.

[14] Mingyang Lu, Zhenjiang Fan, Bin Xu, Lujun Chen, Xiao Zheng, Jundong Li, Taieb Znati, Qi Mi and Jingting Jiang, "Using machine learning to predict ovarian cancer," *International Journal of Medical Informatics*, 2020.

[15] Yasser EL-Manzalawy, Tsung-Yu Hsieh, Manu Shivakumar, Dokyoon Kim and Vasant Honavar, "Min-redundancy and max-relevance multiview feature selection for predicting," *BMC Medical Genomics*, 2017.

[16] Pijush Das, Anirban Roychowdhury, Subhadeep Das, Susanta Roychoudhury and Sucheta Tripathy, "sigFeature: Novel Significant Feature Selection Method for Classification of Gene Expression Data Using Support Vector Machine and t Statistic," *frontiers in Genetics*, 2020.

[17] Zengqi Yue, Chen Sun, Fengye Chen, Yuqing Zhang, Weijie Xu, Sahar Shabbir, Long Zou, Weiguo Lu, Wei Wang, Zhenwei Xie, Lanyun Zhou, Yan Lu And Jin Yu, "Machine learning-based LIBS spectrum analysis of human blood plasma allows ovarian cancer diagnosis," *Biomedical Optics Express*, 2021.

[18] H. Li and O. Chutatape, "Automated Feature Extraction in Color Retinal," *IEEE Transactions On Biomedical Engineering, ,* vol. 51, 2004.

https://digitalcommons.aaru.edu.jo/fcij/vol6/iss2/1
DOI: http://doi.org/10.54623/fue.fcij.6.2.1

[19] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuro-science.*, vol. 3, pp. 70-86, 1991.

[20] R. S and M. R, "An Automatic Bone Disorder Classifiation Using Hybrid Texture Feature Extraction With Bone Mineral Density," *Asian Pacific Journal of Cancer Prevention*, vol. 19, 2018.

[21] U. R. Acharya, S. L. Fernandes, J. E. WeiKoh and E. J. Ciaccio, "Automated Detection of Alzheimer' s Disease Using Brain MRI- A Study with Various Feature Extraction Techniques," *Journal of Medical Systems* , 2019.

[22] A. M. Idrees, A. I. ElSeddawy and M. O. Zeidan, "Knowledge Discovery based Framework for Enhancing the House of Quality," *International Journal of Advanced Computer Science and Applications (IJACSA),* vol. 10, no. 7, pp. 324-331, 2019.

[23] A. M. Idrees and W. H. Gomaa, "A Proposed Method for Minimizing Mining Tasks' Data Dimensionality," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 2, 2020.

[24] M. Lu, Z. Fand, B. Xu, L. Chen, X. Zheng, J. Li, T. Znati, Q. Mi and J. Jiang, "Using machine learning to predict ovarian cancer," *International Journal of Medical Informatics*, vol. 141, p. 104195, 2019.

[25] A. E. Khedr, A. I. El Seddawy and A. M. Idrees, "Performance Tuning of K-Mean Clustering Algorithm a Step towards Efficient DSS," *International Journal of Innovative Research in Computer Science & Technology (IJIRCST)*, vol. 2, no. 6, pp. 111-118, 2014.

[26] A. E. Khedr, A. M. Idrees and A. I. El Seddawy, "Enhancing Iterative Dichotomiser 3 algorithm for classification decision tree," *WIREs Data Mining Knowledge Discovery,* vol. 6, p. 70–79, 2016.

[27] A. M. Idrees and F. K. Alsherif, "A Collaborative Evaluation Metrics Approach for Classification Algorithms," *Journal of Southwest Jiaotong University,* vol. 55, no. 1, pp. 1-14, 2020.

[28] A. E. Khedr, "Business Intelligence framework to support Chronic Liver Disease Treatment," *International Journal of Computers & Technology,* vol. 4, no. 2, pp. 307-312, 2013.

Published by Arab Journals Platform, 2021