ReliefF and Association Rule Mining to Determine Cervical Cancer Causes

Zahraa Naser Shah weli

College of Law, Al Nahrain University, Baghdad, Iraq, .Email:zah2005muh@gmail.com

A R T I C L E   I N F O

A B S T R A C T

Cancer patients till this day suffer from the inability of science to predict the causes of the disease before it occurs. One of the cancers that occupy the minds of many women is the cervical cancer because of the delay in its diagnosis as a result of its multiple and unclear causes, so scientists and researchers need to search for the most causative factors. Machine learning approaches have become one of the best and fastest ways to find associations between symptoms and causes of disease. The use of association rule mining (AR) is very effective if diagnostic features are set up. In this work, feature selection (FS) algorithm named ReliefF is used to reach the most correlated factor, then the Apriori algorithm has been updated to reduce the time and space used, and detects features that are closely related to the class attribute to access most factors that cause cervical cancer. The experimental results of the proposed work indicate a number of cervical cancer risk factors that when combined, indicate a woman's likelihood of developing cervical cancer, which is: the number of years of hormonal contraception is greater than or equal to 15, having any type of cancer or HPV or syphilis or HIV, the number of IUD insertion years exceeded 10, First sexual intercourse smaller than 13 and Number of sexual partners greater than 5. The outcomes of this work help both doctors and women to prevent cancer.

1. Introduction

Every year, enormous treatment information is collected by social healthcare institutions. The primary goal of any dataset is how to discern the correlative features that lead to the results [1]. For the greater part datasets, subset features is relevant, i.e. effective over deciding the endpoint value. Those remaining unimportant features, which would infrequently recognizable from the earlier in true problems, are not educational yet help the general dimensional of the issue space. This increment the trouble and computational trouble set for modeling techniques [2]. Feature determination or feature selection (FS) Might defining as characterized concerning features or relevant features and discarding unimportant ones. One of the vast majority known of FS methods is the reliefF algorithm [3]. reliefF algorithm filters the data used to prepare it for an important stage in this work which is to use one of the simplest AR algorithm that is apriori algorithm to find the most related features in huge dataset. Apriori algorithm suggested by Agrawal in 1993, which is used in data mining (DM) for discovering the possible relationships between data (Data Attributes) through test communicated items.

∗Corresponding author : Zahraa Naser Shah weli

Email addresses : zah2005muh@gmail.com

Communicated by: *Dr. Rana Jumaa Surayh aljanabi.*

The main measures in AR are the support and confidence [4]. Important rules are produced through support and confidence existing in association rule mining [5]. Predicting cancer is an excellent step towards its treatment and prevention, so science is currently moving to link the causes of organic and inorganic cancers to predict it before infection. Cervical cancer has many causes, and it usually occurs as a result of a number of interrelated factors. Find these associated rules in predicting risk causes for cervical cancer is very important for patients and doctors.

Between vast majority cancer infect women is the cervical cancer where 90% of it occurs in low and middle income countries. There are three steps to reduce mortality with cervical cancer are:
1. Prevention.
2. Prediction or early diagnosis.
3. Programmed treatment [6].

Cervical cancer has multiple causes, and it is believed that the most important and most common of these causes is infection with the human papillomavirus (HPV). Among the disadvantages of cervical cancer is the inability to detect it early, which makes it difficult to treat. Another disadvantage is that it infects young women [7]. Fig.1 clarify cervical cancer site.

The best prevention method is to predict the disease before it occurs or to predict its causes [8], which is what researchers are trying to do, and it is the main problem that this work is trying to solve through the use of modified apriori algorithm. The modified apriori algorithm is used to find the relationship between features of cervical cancer to find risk factors for this cancer.

The rest of the work divided into 4 sections: some of the previous researches, detailed explanation of the algorithms used, results of each algorithm applied then the conclusion of the work are presented.
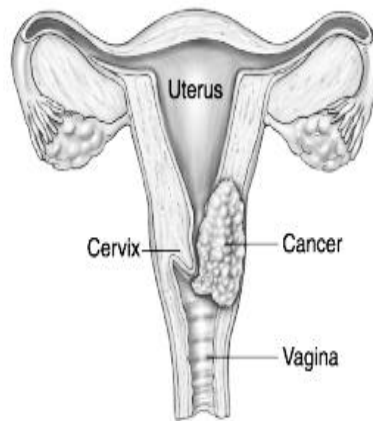

Figure 1 - Cervical cancer site.

2. Related Work

In recent years, many and many researchers have been looking for ways to help people predict the causes of cancer or any disease before it occurs, including data mining methods. An important method of prediction is to use association rule mining that produces certain rules for features that are related to each other. AR mining has various types, such as Tertius, Apriori, FP growth, Predictive apriori and Eclat. In 2011, Jesmin Nahar et al. [9] proposed three AR algorithms, Tertius, Predictive apriori and Apriori algorithms to find factors that prevent some types of cancer such as prostate, breast, lung, cervical, skin and bladder. From the results' table, the best algorithm for this task is apriori algorithm. The rules extracted represent the preventive factors for each type of cancer as the data set used was collected from several sources.

In other work, Y. M. S. Al-Wesabi et al. [10] compares more than one data mining (DM) algorithm where the decision tree gave the best classification result. Two types of FS algorithm, filter and wrapper, are used to discover the best features in the classification. The cervical cancer datasets from UCI that used in this work. The results showed that features (STDs:genital herpes, smokes, number of pregnancies, hormonal contraceptives, first sexual intercourse and age) which gave an accuracy of 97.5%, were these features most closely associated with cervical cancer.

In 2019, Jiayi Lu et al. [11] based on a gene sequence module using data correction, an ensemble strategy, and an assistant gene module the cervical cancer prediction system has been proposed. The risk factor for cervical cancer is examined using performance measures such as recall, precision, accuracy, and F1 score that are used to compare classification algorithms. The proposed model divided into three steps: data correction, an ensemble strategy, and

an assistant gene module. The dataset employed was split into two parts, private dataset from a Chinese hospital and public dataset from Hospital Universitario de Caracas.

In 2020, K. Logeswaran et al. [12] used cervical cancer dataset from UCI repository for prediction. AR mining using Test Train Approach (VARMTTA) and apriori algorithm are employed. The VARMTTA has the best performance than apriori which it's gave 1760 rules where the apriori gave 2317 rules. The accuracy, sensitivity, specificity, positive predictive rate and negative predictive rate were calculated for each target variable.

In 2021, Meera Tandan et al. [13] proposed COVID-19 Symptom Analysis System using association rules mining algorithms such as apriori, FP growth and Eclat where the apriori algorithm gave the best computational time. The dataset used in this work obtained from Wolfram Data Repository. Through this work, it was concluded that cough, fever, sore throat and pneumonia are the most common symptoms in patients with COVID-19.

The best prevention method is to predict the disease before it occurs or to predict its causes, which is what researchers are trying to do, and it is the main problem that this work is trying to solve through the use of modified apriori algorithm.

3. Background

The algorithms that have been used in this work are summarized below:

3.1. ReliefF Algorithm

ReliefF algorithm, suggested by Kononenko in1994 [14], is one of the most FS methods employed in DM. It is an improvement for the original Relief algorithm [15]. The two algorithms find near-hit and near-miss instances but the difference when relief algorithm uses Euclidean distance for this where reliefF uses Manhattan distance. The steps for ReliefF algorithm are as follows [16]:
1. Set initial weight 0.
2. Select instance randomize.
3. Find the nearest hit (same class).
4. Find the nearest miss (different class).
5. Applied estimate weights and steps are repeated, except for the first.

The pseudo code for reliefF algorithm is listed below:

```
Input: for each training instance a vector of attribute values and the class value
Output: the vector w of estimations of the qualities of attributes.
1. set all weights  w[A] =0.0
2. for I =1 to m do begin // m: training data size
3. randomly select an instance rᵢ
4. find k-nearest hits hⱼ
5. for each class C ≠ class(rᵢ) do
6. from class C find k nearest misses mⱼ(c)
7. for A =1 to a
8. w[A]=w[A]- Σ_{j=1}^{k} diff(a,ri,hj)/(m.k) + Σ_{c≠class ri} ( (p(c)/(1-p(class(ri)))) Σ_{j=1}^{k} diff(a,ri,mj(c)) )/(m.k)
9. End
```

3.2. Association Rule (AR) Mining

In DM, the most important association approaches is the AR mining. AR mining can be used in different approaches for finding the relationships between features in dataset and in decisions support. The primary AR metrics that reflect benefits and strongly of rules are minimum support threshold (minsup) and minimum confidence threshold (mincon)[17].

The task for AR mining is:
-   Detect all features whose supports are greater than minimum support.

- Use previously specified features to find the required rules rely on confidence threshold [18].

One of the most popular learning algorithms is the Apriori algorithm. The rest of the learning algorithms are an update of this algorithm [19].

3.3. Apriori

One of the worthy AR mining is the Apriori algorithm. Apriori algorithm as other AR algorithms depend on finding support and confidence and delete all features does not meet minsup and mincon. The basic steps of apriori algorithm in cancer prevention factors are below:

- creates a list of 1- preventing factor
- Obtain the support count for each preventing factors and delete any prevention factors do not satisfy the minsup.
- Connect the 1- prevention factor together to create the 2-prevention factor.
- The steps taken for the 1-prevention factor candidate prevention factor is repeated for the 2-prevention factor candidate prevention factor.
- Continuing to work with the above steps until all frequent prevention factor are finished

Then the confidence is created for each rule produced from prevention factors [20].

Apriori algorithm has some drawbacks such as: poor performance, space and time complexity. Therefore, modification of apriori algorithm must perform through develop Frequent Pattern Growth method when scans the dataset only twice. Also, partition the data to reduce space.

4. Methodology

This work splits into three stages: preprocessing stages when preparing the dataset, feature selection stage when employing the ReliefF algorithm, then the frequent patterns found using AR mining. Fig.2 summarized proposed system architecture.

The proposed work implemented on a system having 4 GB RAM and 1.8 GHz Intel Cori5-3317 processor, 64-bit operating system and Python 3.8.5 programming language.
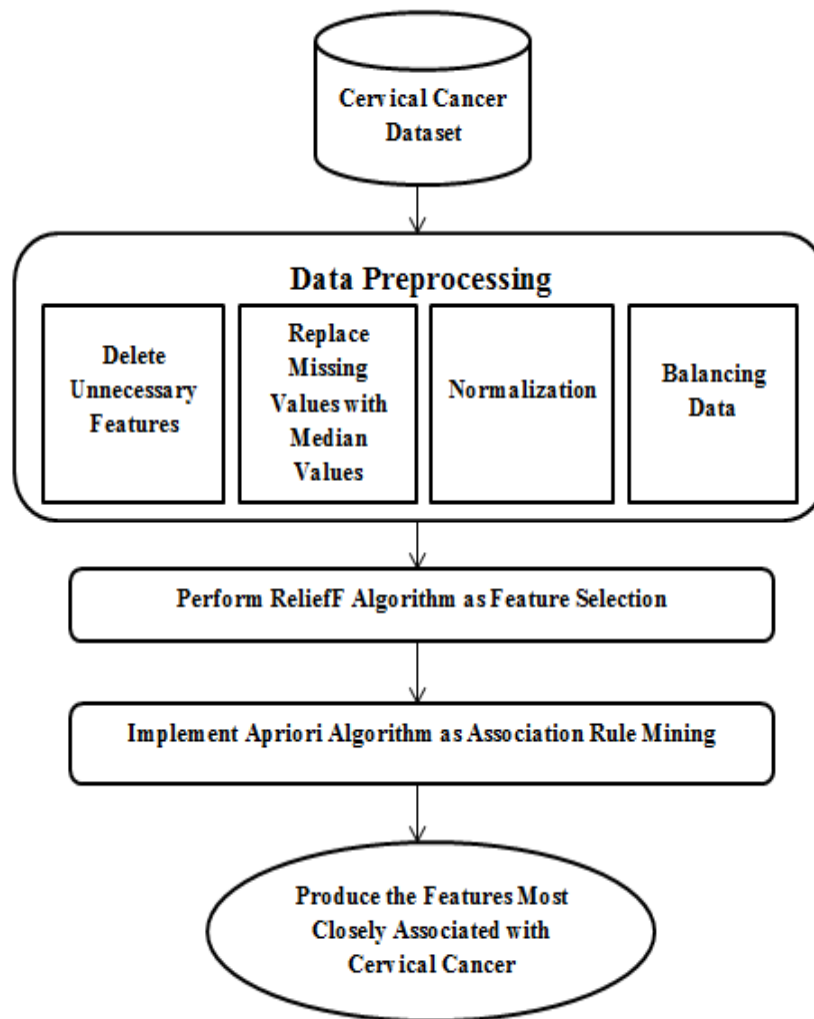
Figure 2. Suggested System Architecture.

The three steps of work are described minutely below:

4.1. Data Preprocessing

The dataset used was collected at Hospital 'Univer-sitario de Caracas' in Caracas, Venezuela [21]. The dataset contains 858 instances (patients) and 32 features (risk factors). The dataset includes medical information, daily habits and demographic information for 858 patients. There are 4 target labels: Biopsy, Schiller, Hinselmann and Cytology. The dataset contains numerous missing values due to patients who did not answer all questions. Table (1) contains the name for 28 features in addition to the four features that were deleted after the first step of the preprocessing.

Data quality is the basis work of DM [22]. Raw data collected in various ways suffers from missing and outlier values as well as noisy data. Therefore, data must preprocess before using to get efficient data. The steps below implemented on used data for preprocessing:

Step 1: delete any feature that contains high ratio of missing value or constant value. Four features are deleted from this dataset, like STDs: Time since first diagnosis, STDs: Time since last diagnosis, STDs:cervical condylomatosis and STDs:AIDS.

Step 2: replace missing values in any features with the median value which computed as in Eq. (1) [23].

Median: x= sort(x),
Median=$x_{n/2}$(Halfbelow,Halfabove)   (1)

Step3: due to benefits of performing normalization on any dataset for the harmony of each feature with other features [24], so performs Min-Max normalization on all remaining features. Min-Max normalization was applied by using Eq. (2) [25].

$$value\ new= \frac{value-min}{max-min}$$ (2)

Step 4: After three steps above, cervical cancer dataset consists of (858) instances and (28) features. Where this data is imbalanced (208 instances are cancerous and 650 are non-cancerous). This cancerous data is when treat 4 target labels: Biopsy, Schiller, Hinselmann and Cytology as one target label. To make the data more balanced was used adaptive synthetic (ADASYN) sampling. The ADASYN method increase minority class sample [26]. The new sample was created as in Eq. (3):

$$Si = Xi + (Xzi − Xi) × λ$$ (3)

X_i: one of the instances
X_{zi}: Randomly minority data example was chosen
$(x_{zi} – x_i)$: the difference vector
$λ$: a random number: $λ ∈ [0, 1]$.

After ADASYN, minority class has oversampled from 208 to 654 instances.

4.2. Feature selection (ReliefF algorithm)

The second step in the proposed work is feature selection based on a ReliefF algorithm. ReliefF algorithm select feature depending on contribution when the distance between features is increased if features have high contribution. The weight in reliefF is between [1, -1], where 1 is for relevant features and -1 for irrelevant features. In this work, reliefF implemented on 28 features remaining after preprocessing step to find the irrelevant features or less correlated features to target class. Any feature has weight less than 0 is ignored. For cervical cancer data sets used in this work, 6 features are ignored. Therefore, 22 features remain in data sets. Table 1explains the weight for each feature after perform reliefF algorithm.

Table 1- The weight for each feature after performing reliefF algorithm

Feature	weight	Feature	weight
Number of sexual partners	0.214	STDs:condylomatosis	0.004
First sexual intercourse	0.1922	Smokes (packs/year)	0.00251
Age	0.118	Dx	0.00224
Num of pregnancies	0.115	Dx:HPV	0.00198
Hormonal Contraceptives	0.104	Dx:Cancer	0.00073
Hormonal Contraceptives (years)	0.0733	STDs:HIV	0.00066
STDs (number)	0.0721	Dx:CIN	0.00031
Smokes	0.0422	STDs:syphilis	0.0003

STDs	0.0401	STDs:vaginal condylomatosis	-0.0005
IUD	0.0271	STDs:genital herpes	-0.002
Smokes (years)	0.0144	STDs:Hepatitis B	-0.0021
STDs: Number of diagnosis	0.011	STDs:pelvic inflammatory disease	-0.003
IUD (years)	0.0058	STDs:molluscum contagiosum	-0.0177
STDs:vulvo-perineal condylomatosis	0.0362	STDs:HPV	-0.0185

4.3. Apriori Algorithm

The last step in this work was implementing Apriori algorithm. The Apriori algorithm used in this work improves the older Apriori algorithm through following points:

1- Minimize the size when compute the frequent 1-itemsets and support for it then compute the frequent 2-itemsets and support for it.

2- Divide the data set into different horizontal partitions by manually selecting some partitions.

Through repeated attempts and experiments, the best minsup threshold used is 0.4 while the mincon is 70%. The pseudo code for proposed apriori algorithm illustrated below:

```
Input: Database D, of transactions;

 Minimum support threshold, min-sup

Output: L, frequent item sets in D

Method:

L(1) =find frequent-1-item sets (D);

For each transaction t belongs to D

     Count- items=count- items (t);

For (k=2; L(k-1)! =null; k++)

{       C(k) =apriori-gen (L(k-1, min-sup) ;

flag=1;
```

```
For each transaction t belong to D

    Where count-items>=k

{    If(flaq==1)

          { c=subset(c (k), t);

               c.count++;

               If (c.count==min-sup) flaq=0; }

       If(flaq==0) Exit from loop;  }

L(k) ={c.count=min-sup}

Return L=U (k) L (k);
```

5. Result and Discussion

After the preprocessing step, 28 features remain. Then after the reliefF algorithm is implemented, 22 relevant features remain. The Apriori algorithm found frequent patterns among these features that representing the most serious causes of cervical cancer and these serious causes are listed below:

- ((Number of sexual partners > 5 & First sexual intercourse< 15 & Hormonal Contraceptives= yes & Hormonal Contraceptives (years) >=15 & STDs: condylomatosis= yes))
- ((Dx:HPV = yes & Smokes (years)≥ 10))
- ((IUD= yes & IUD (years)> 12 & STDs:HIV= yes))
- ((STDs:HIV=yes & IUD (years)>10))
- ((Dx:HPV = yes & Hormonal Contraceptives (years)≥20))
- ((Age ≥50 & Num of pregnancies>6 & Dx:Cancer))
- ((Age >52 & First sexual intercourse <13 & IUD (years)>10))
- ((Age > 45 & Hormonal Contraceptives (years)≥ 19 & Num of pregnancies≥8))
- ((Smokes (years) >18 & Smokes (packs/year)>22 & STDs:syphilis = yes))
- ((Dx:Cancer=yes & STDs (number)≥3))
- ((Number of sexual partners≥7 & First sexual intercourse<20 & IUD (years)≥19 & STDs (number)>2))
- ((STDs (number)>3 & HPV infection=yes & STDs:HIV= yes))

From the results shown above, it is possible to note a several features that, if more than one is get togather, there is a high chance of developing cervical cancer, including: the number of years of hormonal contraception is greater than or equal to 15, having any type of cancer or HPV or syphilis or HIV, the number of IUD insertion years exceeded 10, First sexual intercourse smaller than 13 and Number of sexual partners greater than 5.

The rules extracted using the apriori algorithm gave a useful and meaningful connection to the predisposing features of cervical cancer and a clearer picture for clinicians and people alike. The execution time for modified Apriori algorithm= 16 sec. Where the total execution time for proposed work= 23 second.

In Table (2) a comparison was made between this work and one of the earlier research work No. [27] To explain the contribution:

Table 2- proposed work compare with previous work

	dataset	Algorithms	No. of frequent data patterns	Execution time
Proposed work	UCI cervical cancer risk factors	ReliefF algorithm and Apriori Algorithm	12	23 sec.
Research [27]	UCI cervical cancer risk factors	ANN, Decision Tree, Random Forest, Logistic Model Tree and apriori algorithm	5	Not defined

6. Conclusion

Predicting cancer before occurs or detecting it in its first stages is a very important step for saving humans. Among the disadvantages associated with cervical cancer is that its symptoms appear only in the late stages, so it is necessary to know its causes to predict it and to take the necessary precautions by women. The link between several causes of cervical cancer is the main goal of this work, if it is combined in a woman, are more likely to develop. The data was first initialized, preprocessed, and used reliefF algorithm to select the most influential features and delete the weak features. To link the features with each other, the Apriori algorithm is used after modifying it, which gave a number of logical rules as more dangerous causes of the disease. Therefore, the major approaches for control of cervical cancer involve: prevention of invasive cancer by HPV vaccination and early prediction of common causes in humans which is depends on Number of sexual partners, First sexual intercourse, IUD insertion years and hormonal contraception.

References

[1] A. Batra, M. R and R. Haldar, "Association Rule Mining based on a Modified Apriori Algorithm in Heart Disease Prediction", International Journal of Scientific & Engineering Research, vol. 9, No. 9, (2018).

[2] Rustum Mohammed, H. , "Boundaries Object Detection for Skin Cancer Image using Gray-Level Co-Occurrence Matrix (GLCM) and features points", Journal of Al-Qadisiyah for Computer Science and Mathematics, vol. 7, No. 1, (2017), pp. 160-172. Retrieved from https://qu.edu.iq/journalcm/index.php/journalcm/article/view/94

[3] Ryan J. Urbanowicz, M. Meekerb, W. La Cavaa, R. S. Olsona and J. H. Moorea, "Relief-based feature selection: Introduction and review", Journal of Biomedical Informatics, vol. 85, (2018), pp. 189-203. https://doi.org/10.1016/j.jbi.2018.07.014.

[4] R. Paul, T. Groza, J. Hunter and A. Zankl, "Semantic interestingness measures for discovering association rules in the skeletal dysplasia domain", J Biomed Semantics, vol. 5, (2014). doi: 10.1186/2041-1480-5-8.

[5] D. Azzeddine, S. Jabri, B. Yousse and G. Taoufiq, "The selection of the relevant association rules using the ELECTRE method with multiple criteria", IAES International Journal of Artificial Intelligence (IJ-AI), Vol. 9, No. 4, Des (2020). doi: 10.11591/ijai.v9.i4.pp638-645.

[6] URL:https://www.who.int/health-topics/cervical-cancer#tab=tab_1

[7] E. Ratnasari Putri, A. Vijai Nasrulloh and A. E. Fahrudin, "Coloring of Cervical Cancer's CT Images to Localize Cervical Cancer", International Journal of Electrical and Computer Engineering (IJECE), vol. 5, No. 2, April (2015), pp. 304-310. doi: http://doi.org/10.11591/ijece.v5i2.

[8] Mohammed, Y., & Saleh, E., "Investigating the Applicability of Logistic Regression and Artificial Neural Networks in Predicting Breast Cancer", Journal of Al-Qadisiyah for Computer Science and Mathematics, vol. 12, No. 2, (2020), pp. 63-73. https://doi.org/10.29304/jqcm.2020.12.2.697.

[9] . Nahar, K. S. Tickle, A. B. M. Shawkat Ali and Y. Phoebe Chen, "Significant Cancer Prevention Factor Extraction: An Association Rule Discovery Approach", J Med Syst, vol. 35, (2011), pp. 353–367. doi: 10.1007/s10916-009-9372-8.

[10] Y. M. S. Al-Wesabi, A. Choudhury, D. Won, "Classification of Cervical Cancer Dataset", In: proc. of the 2018 IISE Annual Conference, (2018). doi: 10.13140/RG.2.2.32311.78245.

[11] J. Lu, E. Song, A. Ghoneim and M. Alrashoud, *Machine learning for assisting cervical cancer diagnosis: An ensemble approach*, Future Generation Computer Systems, Elsevier B.V., (2019). doi: 10.1016/j.future.2019.12.033.

[12] K. Logeswaran, P. Suresh, S. Savitha , K. R. Prasanna Kumar, A. P. Ponselvakumar and A. Rajiv Kannan, "Data Driven Diagnosis of Cervical Cancer using Association Rule Mining with Trivial Rule Expulsion Approach", International Journal on Emerging Technologies, Vol.11, No.2,(2020), pp. 110–115.

[13] M. Tandan, Y. Acharya, S. Pokharel and M. Timilsina, "Discovering symptom patterns of COVID-19 patients using association rule mining", Computers in biology and medicine, vol. 131, Feb. (2021). doi: 10.1016/j.compbiomed.2021.104249.

[14] Y. Huang, P. J. McCullagh, N. D. Black، "An optimization of ReliefF for classification in large datasets" ,Data & Knowledge Engineering،Vol. 68, No. 11, (2009), pp. 1348-1356. https://doi.org/10.1016/j.datak.2009.07.011.

[15] D. Jain and V. Singh, "An Efficient Hybrid Feature Selection model for Dimensionality Reduction", In: Proc. International Conference on Computational Intelligence and Data Science (ICCIDS 2018), (2018), pp. 333–341. https://doi.org/10.1016/j.procs.2018.05.188.

[16] Wang, Z. Yan Zhang, Z. Chen, and et al, "Application of ReliefF algorithm to selecting feature sets for classification of high resolution remote sensing image", In: Proc. 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS),   (2016), pp. 755-758. doi: 10.1109/IGARSS.2016.7729190.

[17] A. Meida, D. Palupi Rini and Sukem, "Pattern of E-marketplace Customer Shopping Behavior using Improved Tabu Search and FP-Growth Algorithm", Indonesian Journal of Electrical Engineering and Informatics (IJEEI), Vol. 7, No.4, Des (2019). doi: 10.52549/ijeei.v7i4.1362.

[18] S. SulovaT, "Association Rule Mining for Improvement of IT Project Management", TEM Journal. Vo 7, No. 4, November (2018), pp. 717-722. doi: 10.18421/TEM74-03.

[19] T. Tin Yu and K. Thidar Lynn, "Proposed Method for Modified Apriori Algorithm", In: Proc. Information and Knowledge Engineering | IKE'17, (2017), pp. 68-73.

[20] M. Abdullah Al-Hagery, "Extracting hidden patterns from dates' product data using a machine learning technique", IAES International Journal of Artificial Intelligence (IJ-AI), Vol. 8, No. 3, September (2019), pp. 205–214. doi: 10.11591/ijai.v8.i3.pp205-214.

[21] K. Fernandes, J. S. Cardoso, and J. Fernandes, "Transfer learning with partial observability applied to cervical cancer screening", In: Proc. Iberian Conf. Pattern Recognit. Image Anal., (2017), pp. 243_250.

[22] Adil Raheem, O., & Saleh Alomari, E., "An Adaptive Intrusion Detection System by using Decision Tree", Journal of Al-Qadisiyah for Computer Science and Mathematics, vol. 10, No. 2, (2018), pp. 88 - 96. https://doi.org/10.29304/jqcm.2018.10.2.387.

[23] R. F. Woolson and W. R. Clarke, *Statistical methods for the analysis of biomedical data*, 2nd Edition, John Wiley & Sons, (2002).

[24] Z. Naser Shahweli, "Deep Belief Network for Predicting the Predisposition to Lung Cancer in TP53 Gene", Iraqi Journal of Science, Vol. 61, No. 1, (2020), pp. 171-177. doi: 10.24996/ijs.2020.61.1.19

[25] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance", Applied Soft Computing Journal, vol. 78, (2019).

[26] H. He, Y. Bai, E. A. Garcia and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning", In: Proc. 2008 IEEE International Joint Conference on Neural Networks, Hong Kong, (2008), pp. 1322-1328.

[27] Arif-Ul-Islam, S. H. Ripon, N. Qaisar Bhuiyan, "Cervical cancer risk factors: classification and mining associations", APTIKOM Journal on Computer Science and Information Technologies, Vol. 4, No. 1, (2019), pp. 8-18. doi: 10.11591/APTIKOM.J.CSIT.131.