A proposed Algorithm For Prediction HIV By Using Data Mining Technology

Ihab L. Hussein Alsammak
Ministry of Education, Directorate General of Education of Karbala
ehablaith@gmail.com

ARTICLE INFO
Submission date:30/5/2019
Acceptance date: 27/8/2019
Publication date: 18/9/2019

Keywords: Data mining, Association Rule Mining, HIV, Pattern.

Abstract

Data mining (DM) depends on analyzing wide amounts of information from many sources with a large database to obtain significant information that can be used in different areas, such as education and healthcare, in order to obtain the best possible knowledge. Within the field of healthcare, the huge quantity of medical science information has the potential to predict dangerous illnesses such as human immunodeficiency viruses (HIV) within a short period of time if processed in a suitable way. Therefore, early handling can be set by using appropriate algorithms and methods, such as Association Rule Mining Algorithms, which will affect the diagnosis positively.

The principles of the tree algorithm were based on the construction of our proposed algorithm, This paper proposes an algorithm that depends on a repeated pattern that can help doctors to distinguish the severity level of AIDS in patients by means of a database, in experimental analysis on a data set between the database of patients and the proposed algorithm. The proposed algorithm consistently takes less time to find the people with HIV as compared to the original database ,and a difference in the database memory storage has been found between the database of patients and the proposed algorithm.

Traditional algorithms used in previous, such as PART, J48, and Naïve Bayes, have performed poorly to predict disease, and this problem has been solved by new technologies and the development of a predictive system for HIV status. Thus, the main goal of this research as the final results showed that the proposed algorithm is workable in an early prediction of HIV with less time and less memory storage compare it with the original database and reducing the database to a mini tree that can help predict the condition of the patient at an early stage.

1. INTRODUCTION

During the last two decades over sixty million people ,have been infected with HIV/AIDS. The discovery of HIV/AIDS at an early stage is the key to its treatment. It is one of the most serious causes of death, as it is a disease which causes uncontrolled cell increase. In the 1980s, HIV was first identified. From this period onwards, there was an increase in studies and research into various aspects of the disease. Since 2006, approximately four million people in the world are estimated to have been infected with AIDS. An estimated ten million live with this virus. AIDS is a lent virus, i.e., it targets the immune system with long incubation periods. The disease can also be found in animals.

Many countries suffer from problems in the field of predictive healthcare. DM methods can be applied using pattern recognition to hold classifying cases of AIDS in developing countries. When new information is extracted from a database, in many fields, user databases have such been developed that it enables new information to be held and extracted. This allows organizations to obtain accurate information, query processing and make use of applications as well as carries out decision-making. Within the field of healthcare, such databases enable researchers to carry out data mining, which can help predict the outlook of viruses, in this case HIV/AIDS[1].

Data mining involves a set of parameters, such as a history of the disease and knowledge about its characteristics. It makes use of a set of technologies to find data. Therefore, an automated predictive method is necessary to predict the percentage of the disease [2].

There are many ways that the virus can be transmitted, such as unprotected sex, hypodermic needles, from mothers to babies during pregnancy, contaminated blood transportation, birth or breastfeeding. The virus shortens a person’s life by attacking their immune system, specifically the white blood cells, known as the CD4 cells T cells, cluster of differentiation, is a type of white blood cells that are spread throughout the body, and whose mission is to find and destroy viruses, bacteria, and invasive germs and responsible for fighting the infection. The virus can reduce the number of CD4 cells, which can lead to other infections or cancers. Anti-retroviral handling is one of the top therapies for AIDS patients. This form of treatment can delay the course of the disease [3]. In spite of the variety of techniques used previously, the simplest and most obvious techniques were chosen to obtain accurate and improved results at the same time .Through the study of basic data mining algorithms, especially Association Rule Mining Algorithms, new algorithms can be proposed, which are based on the principles of the old algorithms but optimized. This can be obtained by using search algorithms in repetitive patterns, as well as the use of new databases in the style of tree algorithms. These databases can remove any repetitive data and reduce its amount in the database based on the relationships between elements. This results in an integrated gap-free algorithm.

The paper is organized as follows. **Section 2** presents some of the previous work and related studies that have been carried out in the field. It also highlights other similar research and methods. In **Section 3,** the research problem is clarified and the motivation behind the case study is put forward. **Section 4** introduces the methodology that will be used in this study. **Section 5** describes the proposed algorithm, while **Section 6** presents the experiment results and explains the proposed system with an example. Finally, in **Section 7**, the conclusion is given.

2. LITERATURE REVIEW

Many researchers have studied HIV. In recent years, there has been a particular emphasis on association rule mining and proposed algorithms which can take care of missing values for detecting AIDS. Such studies include Avudaiselvi and Tamizhchelvi, [4] and Clare and Avanija [5] who have used expectation-maximization (EM) algorithms, neural networks, decision trees, naive Bayes and linear regression. Dom, et al. [6], Nahar, Tickle and Chen and Imam used an Apriori algorithm, while Rameshkumar [7] developed a model using association rule mining to find missing values for detecting AIDS. Monard, Prati, and de Carvalho [8] developed a predictive DM approach, using many methods to extract knowledge rules from DB(database). The negative impact on the results is because of the missing values and the loss of time required to search for holders of the disease with the loss of storage space. They also proposed a new approach in association rules to avoid this disadvantage.

The studies of the techniques used in previous research aim to determine the proportion of the probability of AIDS based on large data analyzes. In previous research it collects patterns of medical data to find symptoms of the disease and gaps were found when dealing with large data, especially when using decision tree, support vector machine, which one of its problems not to reduce the databases and the seizure of all elements. Without filtering, a new model and algorithm was proposed to systematically close these gaps if the data extraction techniques are applied based on an algorithm that deals with large ,systematic databases, enabling us to provide areliable performance. The data must be extracted in a specific format so that it can be used for knowledge discovery.

3.DEFINITION OF PROBLEM

AIDS is considered a global pandemic, with almost 35.3 million people living with the disease. Approximately16.8 million sufferers are women, while 17.2 million are men. Furthermore, roughly 3.4 million sufferers are under the age of fifteen. In 2010, there were approximately 1.8 million deaths from AIDS, while in 2005, there were 2.2 million, A total of 800 medical datasets have been collected with many attributes that have Boolean values in addition to age and sex. The input data has been collected from a government data portal for the National AIDS Control Organization (NCBI)[9][10].

Medical science has a large possibility of discovering unclear inputs. Current techniques are tending towards the prediction of future diseases, given that all have clear symptoms. For this reason, this paper will propose to use data mining as a predictive tool and techniques to collect a sizeable amount of health data. This data will be normalized using a normalization technique. In this database, all variables will be analyzed in order to make suitable decisions that control the prediction of AIDS. This will be carried out using the concept of clustering. The size of the data will also be reduced using vertical fragmentation. More specifically, this research will propose a database in which an improved algorithm can be implemented, based on the tree algorithm theory. This algorithm has the potential to be used to filter data in a large database and extract hidden information efficiently [11].

The principles of the tree algorithm were based on the construction of our proposed algorithm, that can help doctors to distinguish the severity level of AIDS in patients by means of a database ,in experimental analysis the proposed algorithm takes consistently less time and database memory storage to find the people with HIV. HIV testing is necessary to slow down the spread of the infection. As many people are not aware of the HIV infection, they may be less likely to take precautions to help prevent the spread of the virus to other people. Early diagnosis often leads to treatment with drugs that may delay the development of AIDS [12][13].

4. APPLIED METHODOLOGY

In order to reduce the size of data used for patient databases, data cleaning and handling must be carried out. This involves accessing the single record of each patient and constraining most probabilities of values by assigning the likelihood of bucket values to them. The approach used in this study is helpful in distinguishing the presence of the virus in a patient. Its end result will enable handling procedures and decisions to be made more carefully by practitioners. While building the database basic need analysis, HIV infection is diagnosed by laboratory tests:

1. The ELISA test is performed after a month of stopping the virus by examining a blood sample and its outcome.
2. Testing the examination of the manifold (Western Pluton), a confirmation test for positive case studies of ELISA.
3. Quick inspection.
4. Antigen examination is conducted two weeks after exposure to infection and its result in the early future group of infection and stages of the disease.
5. PCR is performed after a week of exposure.

Figure1 represents the main operations of the suggested methodology of the current research. There are three phases of the disease that can be identified by knowing the patient's symptoms. After, the level of severity to the patient can be determined through the necessary analysis and diagnosis of the development of the disease. According to appropriate algorithms, a database can also be used to work this out. Traditional algorithms used in previous research such as PART, J48, and Naïve Bayes have performed poorly in predicting disease and have been solved by new technologies and the development of a predictive HIV status system. The general approach of the methodology, which begins with the order of data, its preparation, modeling and configuration of the controls are adopted in the proposed algorithm. These steps are applied to any proposed database, whether large or small, according to the number of patients without restricting it to a certain number.

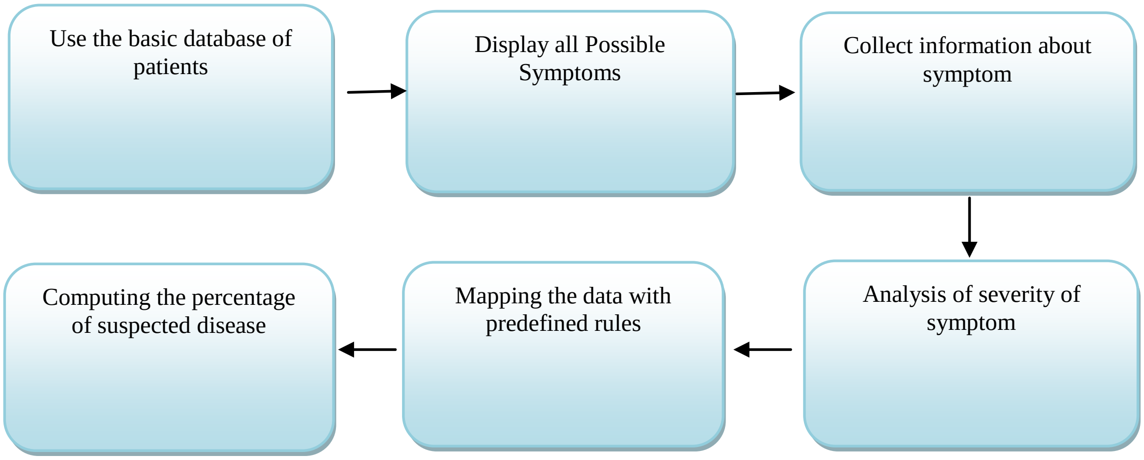

Figure1: Component of Methodology

5. THE PROPOSED ALGORITHM

For this study, a TDB(transaction database) of patients using Min_Sup(minimum support) has been selected. As mentioned above, the disease has three stages and is, therefore, encoded in different ways. This is the first step to building the algorithm.

5.1 Steps of the proposed method:

- Depending on the Min_Sup value, if its support value is less than Min_ Sup, the patients name will be deleted from the database, otherwise the database will be retained by name.
- The symptoms of a patient in database select as SP (symptoms of a patient) determined based on the most common symptoms in the patient.
- Be based on the updated database:

1.Sort the database by symptoms.
2.Create Root as NULL for tree.
3.For each symptoms of a patient in database select as SP and go to 4.
4.Should a descending ordered SP(symptoms of a patient) by [A1 | B1]. A1 is the first symptoms and B1 is have a rest of the symptoms in each SP in database .
5.Check If A1 = the a large amount of FI(frequent item sets), go to 6. else go to 8.
6.If Root in tree has a direct new node as child then move the root to A1.
7.For each item reset in part B1:
generate a new node from old root.
If B has no new node and already exists in tree then not add in tree.
If B1 new root but no new node then add only Bi .

8.do not add this SP in tree .
9.The output new tree with Si values
10. If S0 Min_Sup value=0 then threshold is equal to zero, which did not provide any of the above symptoms
11. If S1 Min_Sup value=1 then threshold equals one ,patient entering the first stage of the disease.
12. If S2 Min_Sup value=2 then threshold equals two ,patient entry into the second stage of the disease.
13. If S3 Min_Sup value=3 then threshold is equal to three patients entering the third stage of the disease. F is selected as a frequent item.

If the patient has symptoms of the disease but not a common symptom of the disease is excluded from the database, in case the patient suffers from minor symptoms but within the suspected symptoms are entered into the algorithm and depending on these symptoms and severity is the final tree is drawn according to the repeated numbers at each injury determines the level of disease. The method can be followed in the flowchart in figure 2.

Figure2:Details of the Proposed Algorithm

6. Implementation of an Algorithm with an Example

Initially, this database was proposed to illustrate the suggested algorithm in detail. This algorithm was applied to a large database of 60,000 patients and once again to 100,000 patients with different symptoms. A proposed database of only twenty cases will be used to illustrate the work of the database with the algorithm. The principle of proposed algorithm depends on the tree algorithm. In implementation, the same database presented in Table1 will be used and the support for each item will be calculated.

Table 1: Transactions of Database

Case NO.	TDB
Case1	Headache, diarrhea, nausea and vomiting
Case2	Headache, long-lasting diarrhea, nausea and vomiting, fatigue, aching muscles, sore throat, swollen lymph nodes, a red rash that does not itch (usually on torso), fever that lasts for more than ten days, being tired all the time, swollen lymph nodes in neck or groin, night sweats, unexplained weight loss, purplish spots on skin that do not go away, severe shortness of breath, ,yeast infections in mouth, throat or vagina, bruises or bleeding that cannot be explained
Case3	Nausea and vomiting, fatigue, aching muscles, sore throat, swollen lymph nodes, a red rash that does not itch (usually on torso)
Case4	Diarrhea, nausea and vomiting
Case5	Swollen lymph nodes, headache
Case6	Sore throat, swollen lymph nodes, a red rash that does not itch (usually on torso), fever
Case7	Fatigue, aching muscles, sore throat, swollen lymph nodes
Case8	Diarrhea, nausea and vomiting
Case9	Headache, diarrhea, nausea and vomiting, fatigue, aching muscles
Case10	Headache, long-lasting diarrhea, nausea and vomiting, fatigue, aching muscles, sore throat, swollen lymph nodes in neck or groin, a red rash that does not itch (usually on torso), fever that lasts for more than ten days, being tired all of the time, S night sweats, unexplained weight loss, purplish spots on skin that do not go away, severe shortness of breath, yeast infections in mouth, throat or vagina, bruises or bleeding that cannot be explained
Case11	A red rash that does not itch (usually on torso), fever
Case12	A red rash that does not itch (usually on torso), fever
Case13	Headache, diarrhea, nausea and vomiting, fatigue, aching muscles ,fever, sore throat
Case14	Diarrhea, nausea and vomiting
Case15	Headache, diarrhea, nausea and vomiting, fatigue, aching muscles, sore throat, swollen lymph nodes, a red rash that does not itch (usually on torso), fever
Case16	Swollen lymph nodes, a red rash that does not itch (usually on torso), fever, sore throat, diarrhea, nausea and vomiting.
Case17	Headache, diarrhea, nausea and vomiting, fatigue
Case18	Headache, diarrhea, nausea and vomiting, fatigue, sore throat, swollen lymph nodes, a red rash that does not itch (usually on torso)
Case19	Fatigue, aching muscles
Case 20	headache, long-lasting diarrhea, nausea and vomiting, fatigue, aching muscles, sore throat, swollen lymph nodes, being tired all the time, swollen lymph nodes in neck or groin, fever that lasts for more than ten days, night sweats, unexplained weight loss, purplish spots on skin that do not go away, severe shortness of breath, , yeast infections in mouth, throat or vagina, bruises or bleeding that cannot be explained

The first phase of the algorithm involved examining the database to confirm the symptoms of the patient. Then, the necessary analysis was conducted in the event of passing the initial phase of the symptoms, which will be referred to as the symbol S1 in Table 2.

This result will require the patient to add the result of the required analysis in addition to the symptoms to find out the level of S in the database, depending on the second table. The algorithm is applied to produce Table 2, which demonstrates the Min_Sup values. It is not a condition that the symptoms should determine the level of the disease, one should rely on the required medical tests described in advance to confirm the stages of the disease.

Table 2: Cases With Min_Sup. Threshold

Case NO.	Min_Sup.
Case1	**S0**
Case2	**S3**
Case3	**S1**
Case4	**S0**
Case5	**S0**
Case6	**S1**
Case7	**S١**
Case8	**S0**
Case9	**S1**
Case10	**S3**
Case11	**S0**
Case12	**S0**
Case13	**S2**
Case14	**S0**
Case15	**S2**
Case16	**S2**
Case17	**S0**
Case18	**S1**
Case19	**S0**
Case20	**S3**

After applying the first and second step of the algorithm, all cases that carry the value of S0 were deleted, which resulted in Table 3.

Table 3: Cases After Algorithm Applied

Case No.	Min_Sup. Threshold **(S)**
Case2	**S3**
Case3	**S1**
Case6	**S1**
Case7	**S1**
Case9	**S1**
Case10	**S3**
Case13	**S2**
Case15	**S2**
Case16	**S2**
Case18	**S1**
Case20	**S3**

For an easier application of the algorithm, each of the symptoms will be referred to with a code as shown, in Table 4.

Table 4: The Symptoms with a Code

Code	Symptoms
A1	Headache
B1	Diarrhea
C1	Nausea and vomiting
D1	Fatigue
E1	Aching muscles
A2	Sore throat
B2	Swollen lymph nodes
C2	Red rash that does not itch (usually on torso)
D2	Fever
E2	CD4 T< 200
A11	Being tired all the time
B11	Swollen lymph nodes in neck or groin
C11	Fever that lasts for more than ten days
D11	Night sweats
E11	Unexplained weight loss
A22	Purplish spots on skin that do not go away
B22	Severe shortness of breath
C22	Long-lasting diarrhea
D22	Yeast infections in mouth, throat or vagina
E22	Bruises or bleeding that cannot be explained

After applying the algorithm steps, the final tree of the proposed database is explained in Figure3. To see all the important symptoms, the final tree was created and to find out links between those symptoms of a new node for each symptoms, the main goal is to extract the tree as a small tree branches where they delete the excess branches, according to the proposed algorithm, find short way which has found the disease. Through the final tree ,the correlation of the symptoms of the disease can be seen ,as well as which the most common symptoms. These should be a branch of the tree to determine the level of disease.

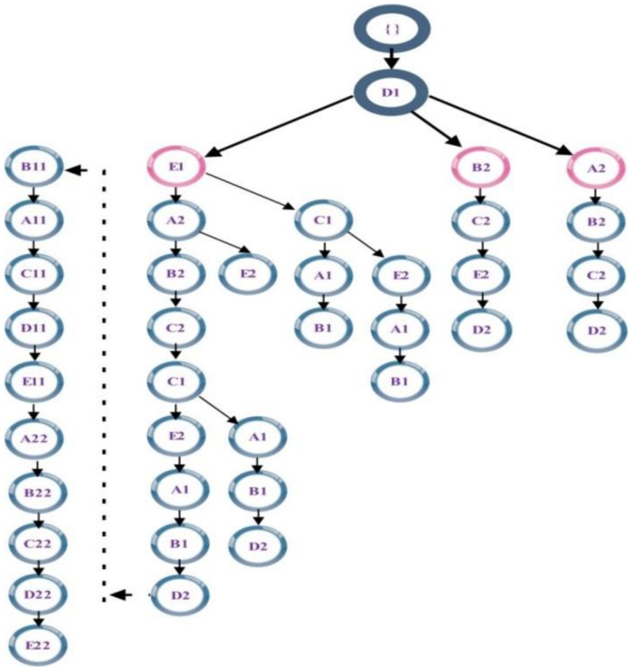

Figure 3: The Final Tree

Through the final tree, the proposed algorithm can be implemented to determine the stages of the disease for each patient, as represented in Figure 4.

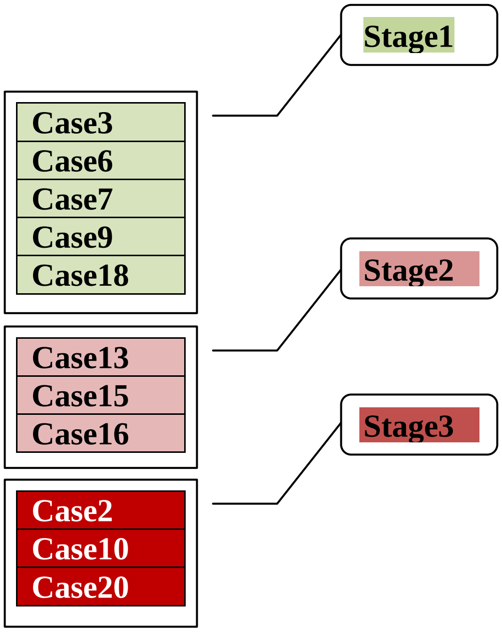

Figure 4: Stages of Disease

In order for the proposed algorithm to be successful in the required form in addition to the prediction of the disease must take into account two important points, namely time and space. In an experimental analysis on a data set containing 60000 patient with 10 symptoms, a comparison has been made between the database of patients and the proposed algorithm. The implementation of the search was done using Net Beans IDE 7.4. ,and the running of the proposed algorithm was obtained in java language. The proposed algorithm takes consistently less time to find the people with HIV as compared to original database. Five readabilities have been assumed, each with a reading different from the other symptoms, as explained in Figure 5 .

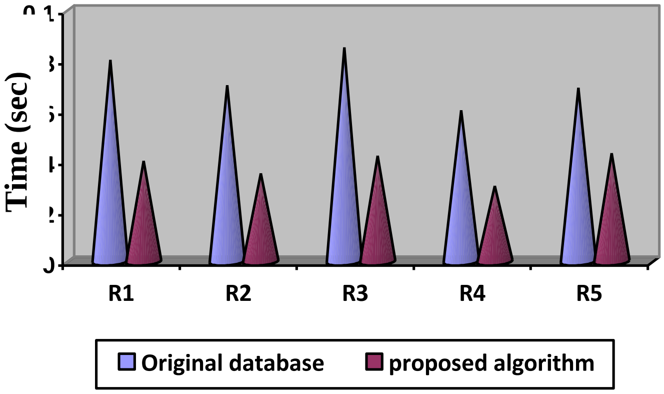

Figure 5: Comparison of the necessary search time between the original database and the proposed algorithm database

After having conducted experiments , a difference in the memory storage is noticed between the database of patients and the proposed algorithm, in the second experimental analysis the original database with 100000 Patients showed a different value of Min_Sup . Six readabilities have been considered, each different from another in symptoms. The proposed algorithm takes consistently less memory storage, as shown in Figure6.

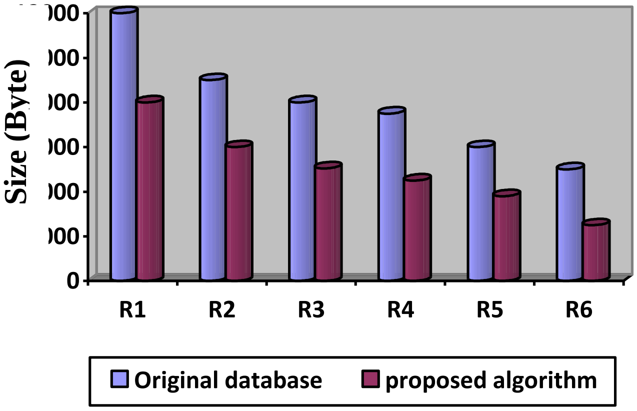

Figure 6: Comparison of the necessary memory storage between the original database and the proposed algorithm database

7.CONCLUSION

Many countries suffer from problems in the field of predictive healthcare. The extraction of new information from a database helps organizations to obtain accurate information, query processing, make use of applications and carry out decision-making. For this reason, data mining must be used to predict viruses. The main goal of this research has been to predict the viruses, as the final results showed that the proposed algorithm is workable in an early prediction of HIV with comparatively less time and less memory storage. The final results have shown that the proposed algorithm is workable in an early prediction of HIV.

If processed in a suitable way, large amounts of medical science information can be used to predict dangerous illnesses, such as HIV, in a shorter period of time. This can be achieved through early handling by means of an algorithm. In this study, the proposed algorithm depends on the tree algorithm and repeated patterns, which can help doctors to distinguish the severity level of AIDS in patients. These databases can remove any repetitive data and reduce the amount of data in the database based on the relationships between elements. This results in an integrated gap-free algorithm.

This can be achieved by using a database and suggesting an algorithm based on many connected symptoms and their level of severity. All the important related attributes of the disease are selected from the dataset in order to discover HIV at an earlier stage. The study has also focused on distinguishing hidden information from a large dataset. Through the suggested example and the results of applying the algorithm to the proposed data, the main objective of the study has been achieved: reducing the database to a mini tree that can help predict the condition of the patient at an early stage.

ACKNOWLEDGMENTS

The authors would like to express their thanks and gratitude to the Head of the Department of Transitional Diseases in the holy city of Karbala, as well as **Dr. Layth Hesson Mirzha**, for his ability to help us to provide full support from the medical side to identify the symptoms of AIDS.

CONFLICT OF INTERESTS

There are no conflicts of interest.

References

[1] Ling Yuan, Wasan Itwee, and Shikang Wei, , Association Rule Mining Technique with Optimized FP-Tree Algorithm, IJASCST 6 issue 5, 2017.

[2] G. Seetha Lakshmi, , A Comparative Study of Data Mining Classification Technique Using HIV/AIDS and STD data, An ISO 3297 Vol. 5, Special Issue 1, March 2017.

[3] Cerna, D, P., and Abdulahi, J, T., "Prediction Of Anti-Retovial Drug Consumption for HIV Patient in Hospital Pharmacy Using Data Mining Technique" International Journal Information Technology and Computer Science, Vol. 2, pp: 52-59, 2016.

[4] Avudaiselvi, S., and Tamizhchelvi, P., "HIV Nomogram Using Big Data Analytics" International Journal of Computational Science and Information Technology, , Vol. 4, Issue. 1, pp: 83-88, February 2016.

[5] Clare, D., and Avanija, J., "Contextual Learning Approach to Impove Diagnostic Accuracy for Hybrid (Lung, HIV and Heat) Diseases", IJSTE International Journal of Science Technology & Engineering, Vol. 2, Issue. 11, pp: 585-588, May 2016,.

[6] Dom, M., *et al.*, , “The Prediction of AIDS Survival: A Data Mining Approach” International Conference on Multivariate Analysis and its Application in Science and Engineering, Vol. 2, pp: 48-53, 2016.

[7] Jesmin Nahar, Tasadduq Imam, Kevin S. Tickle, Yi-Ping Phoebe Chen. , “Association rule mining to detect factors which contribute to heart disease in male and females” Expert Systems and Applications 40, pp. 1086-1093, 2013

[8] Mariammal, D., Jayanthi, S., and Patra, K, S, P., “Major Disease Diagnosis and Treatment Suggestion System Using Data Mining Techniques” International Journal of Advanced Research in Computer Science & Technology, Vol. 2, Issue. 1, pp: 338-341.

[9] Adebayo, P, I., *et al.*, , “The Prediction of Paediatric HIV/AIDS Patient Survival: A Data Mining Approach” Asian Journal of Computer and Information Systems, June 2016, Vol. 4, Issue. 3, pp: 88-94, March 2014.

[10] Ahmad LG, Eshlagby A T, Poorebrabimi A, “Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence” Health and Medical Informatics vol.04 issue 02 1000124. Open access journal, 2013.

[11] Gitanjali J,C. Ranichandra, M. Pounambal “APRIORI algorithm based medical data mining for frequent disease identification” IPASJ International Journal of Information Technology (IIJIT) Vol.2, Issue 4, , ISSN 2321-5976, April 2014.

[12] Priyanka Sharma, DBV Singh, Manoj Kumar Bandil, Nidhi Mishra “Decision Support System for Malaria and Dengue Disease Diagnosis (DSSMD) “International Journal of Information and Computation Technology ISSN0974-2239, Vol.3, Number 7 pp.633-640, 2013.

[13] Adebayo, P, I., *et al.*, “The Prediction of Paediatric HIV/AIDS Patient Survival: A Data Mining Approach” Asian Journal of Computer and Information Systems, Vol. 4, Issue. 3, pp: 88-94, June 2016.

خوارزمية مقترحة للتنبؤ بفيروس نقص المناعة البشرية عن طريق استخدام تكنولوجيا تنقيب البيانات

يعتمد تنقيب البيانات (DM) على تحليل كميات كبيرة من المعلومات في قواعد البيانات الكبيرة بالاعتماد على العديد من المصادر للحصول على معلومات مفيدة يمكن استخدامها في مجالات مختلفة، كالتعليم والرعاية الصحية ، للحصول على أفضل معرفة ممكنة. في مجال الرعاية الصحية، فإن الكم الهائل من معلومات العلوم الطبية لديها القدرة على التنبؤ بالأمراض الخطيرة مثل مرض العوز المناعي البشري (HIV) في فترة زمنية قصيرة إذا تم معالجتها بطريقة مناسبة. لذلك، يمكن ضبط المعالجة المبكرة باستخدام الخوارزميات والأساليب المناسبة، مثل خوارزميات تعدين قواعد البيانات، والتي سوف تؤثر بشكل إيجابي على التشخيص.

تم الاعتماد على مبادئ خوارزمية الشجرة لبناء الخوارزمية المقترحة ، الخوارزمية المقترحة في هذه الورقة تعتمد على الأنماط المتكررة والتي يمكن أن تساعد الأطباء في تمييز مستوى شدة مرض الإيدز لدى المصابين بالاعتماد على قاعدة البيانات، في التحليل التجريبي للخوارزمية المقترحة على مجموعة معلومات قاعدة بيانات المرضى، تستغرق الخوارزمية المقترحة وقتًا أقل للعثور على المصابين بفيروس نقص المناعة البشرية مقارنةً بقاعدة البيانات الأصلية ونجد اختلافًا في ذاكرة تخزين قاعدة البيانات بين قاعدة بيانات المرضى والخوارزمية المقترحة.

لقد ادت الخوارزميات التقليدية المستخدمة في البحوث السابقة مثل PART, J48, and Naïve Bayes اداء ضعيفا للتنبؤ بالأمراض وتم حل هذه المشكلة عن طريق التقنيات الجديدة وتطوير نظام تنبؤي لحاله فيروس نقص المناعة البشرية. وبالتالي، فإن الهدف الرئيسي من هذا البحث كما أظهرته النتائج النهائية أن الخوارزمية المقترحة قابلة للتطبيق في التنبؤ المبكر لفيروس نقص المناعة البشرية مع وقت أقل وذاكرة خزنيه اقل مقارنه مع قاعدة البيانات الاصلية و تقليل قاعدة البيانات إلى شجرة صغيرة يمكن أن تساعد في التنبؤ بحالة المريض في مرحلة مبكرة في اقل وقت ممكن.

الكلمات الدالة: تنقيب البيانات، قواعد الارتباط، مرض النقص المناعي، الانماط المتكررة.