Extraction of Association Rules from Cancer Patient's Records using F-P Growth Algorithm

Razan Alharith[1], Mohammed Khalil [2], Ashraf Osman Ibrahim[3, *] and Salih Hassan Babiker[4]

[1] School of computing and artificial intelligence, Southwest Jiaotong University, Chengdu, China
[2] Faculty of Mathematical Sciences, University of Khartoum, Khartoum, Sudan.
[3] Creative Advanced Machine Intelligence Research Centre, Faculty of Computing and Informatics, University Malaysia Sabah Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia
[4] Faculty of Computer Science and Information Technology, Alzaiem Alazhari University, Khartoum North 13311, Sudan.

*Correspondance : Ashraf Osman Ibrahim ; (ashrafosman@ums.edu.my)

ABSTRACT

Cancer is a leading cause of mortality worldwide, and Sudan has a high cancer burden. The issue is that the data acquired from cancer patients grows yearly, and standard methodologies for analyzing this data are no longer adequate. Data mining techniques such as frequent pattern analysis and association rule mining are utilized in this research to assist in identifying hidden patterns and relationships in data. These strategies were utilized to provide valuable insights into the spread of cancer in Sudan and to assist healthcare professionals in making better diagnosis and treatment decisions. Support and confidence were utilized as measurement criteria. Support is used to evaluate the frequency of occurrence of an item or set of items among all transactions. In contrast, confidence is used to assess the strength of the relationship between groups of things. According to the findings, women are more likely than men to be diagnosed with cancer. The most common cancers in both genders include breast, prostate, ovarian, esophagus, and cervical cancers.

Keywords: Cancer, Association Rules, F-P growth, Confidence, Data Mining.

1 INTRODUCTION

Cancer describes several diseases that could affect distinct organs or any part of the body. Ordinarily, cells increase and decrease in an organized manner. They are assisting us in developing and changing tired tissue and recovering from injuries. Sometimes, something goes wrong, and the cells grow out of control [1]. Cancer, or solid tumors, generally refers to cancers that arise from epithelial surfaces and cells that line glands [2]:

Kidney
Ovary
Liver
Skin
Intestines
Breast
Testis
Adrenals
Cervix
Glands
Bronchus
Prostate

The wide variety of most cancer instances improved and became more tragic in Sudan, as health authorities in Sudan revealed that 40,069 cases of cancer appeared during the past four years at Khartoum Oncology Hospital, similar to a lack of expertise to determine the root reasons for the enormous annual quantity of most cancers instances and assist the Ministry of Health in figuring out the source of cancer. There are numerous difficulties that policymakers experience in understanding the actual scale of the advanced and future cancer troubles. In some areas in which a particular type of cancer is distributed, the affected person's age and gender are factors that play a vital role in making decisions that can help lessen the prevalence of most cancers. So, we investigate the data mining technique (association's rules) to analyze statistics to provide policymakers with an accurate result that effectively displays the proper problem and its size, assisting decision-making.

Data mining and knowledge discovery techniques aim to efficiently locate the most applicable and exciting patterns and characteristics from the specified dataset [3]. Data mining can be applied to different types of databases and data repositories. However, the particular patterns that are found rely on the data mining functionality and techniques used, including prediction, association, correlation analysis, classification, cluster analysis, and class or concept definition [4].

The objectives of this paper concentrated on investigating cancer data in Sudan. Here, we can divide these objectives into:

1) To study the incidence and impact of cancer in Sudan. This involves understanding the prevalence and types of cancer in the country and assessing the impact of cancer on the population.

2) To explore traditional methods for analyzing cancer data and assess their performance and limitations. This objective involves studying the existing literature review and identifying gaps in the current methods.

3) To propose a new method to offer more accurate results for classifying interesting relationships between cancer-related factors.

By achieving the above objectives, this study aims to contribute to the field of cancer research in Sudan and increase the management and understanding of cancer.

The contribution of this paper is to implement data mining approaches, frequent pattern analysis, and association rules in the analysis of cancer data in Sudan. These approaches aim to uncover hidden patterns and relationships in the data that traditional approaches cannot reveal. In doing so, the study provides a valuable understanding of the spread of cancer in Sudan; this practice can support medical practitioners in making more informed decisions regarding diagnosis and treatment.

The paper structure will be as follows: Section two summarizes the work conducted using data mining to analyze data. While Section Three describes the study's methodology, Section Four presents the results and discussion. Finally, section five contains a conclusion and recommendations for future work.

2 RELATED WORKS

Many recent research studies have discussed using data mining in the health sector. Some try to adopt many classification and clustering techniques to predict diagnoses based on some symptoms, and others use association rule mining techniques to generate rules that help them learn more about the exciting relationships between factors and how they correlate to each other.

This paper uses FP growth algorithms as a better association rule mining algorithm approved in previous works. It has high performance in generating rules with minimum execution time and primary memory usage, regardless of whether you use the Apriori, Eclat, or Frequent Items algorithms, which consume much time and memory to generate rules.

According to Global Cancer (GLOBOCAN), there were an expected 18.1 million new cancer cases and 9.6 million cancer deaths in 2018. In both genders, lung cancer was the most well-known cancer (11.6%). Breast cancer in women is the most cancerous (11.6%), followed by prostate cancer (7.1%), colorectal cancer (6.1%), colorectal cancer prevalence (9.2%), most stomach cancers (8.2%), and liver cancer (8.2%) [5]. As stated in the referenced research [10], Breast cancer is the most widespread type of cancer in women, with an incidence rate of 3.9 cases per 100,000 women. The most prevalent histological diagnosis among the 4,423 breast cancer patients was invasive breast carcinoma of no distinct type (NST), accounting for 79.5% (3,517/4,423) of the total cases. The spatial analysis identified regions in Sudan, specifically the States of Nile River, Northern, Red Sea, White Nile, Northern, and Southern Kurdufan, that had a high risk of breast cancer.

The proposed approach to extracting correlation rules from medical records using multi-criteria decision analysis. The primary purpose of this study is to determine the relationship between diseases, symptoms, diseases, and medications. Three methods, which are Apriori, FP Growth, Eclat, Apriori TID, and RElim, will be used, and the closing algorithms will be used to define the correlation rules. After the correlation rules are created with these algorithms, a multi-criteria decision analysis is conducted to select the best algorithm. Algorithms are evaluated based on three limits: implementation, memory space, and reaction time. After performing the examination, the best algorithms were found to be RElim, Apriori TID, and FP-Growth, and all of these algorithms generated the same rules; these algorithms use the lever, the correlation scale, to determine the best rules [6].

The paper provides a concise overview of lung cancer, including its symptoms and several staging strategies. Ultimately, they determine that the Nave Bayes model is the most efficient for forecasting lung cancer patients, followed by IF-THEN, Decision Trees, and lastly the Neural Network. The outputs of decision trees are more readily comprehensible and interpretable. Naïve Bayes outperformed decision trees due to its ability to accurately detect all medical factors. Understanding the correlation between features generated by a neural network is challenging [7].

Highlighted in this case [8], frequent item mining is recognised as a crucial field within pattern mining due to its ability to facilitate effective classification, clustering, and predictive analysis. The algorithms employed in this paper include Apriori, FP Growth, and Eclat. They are utilised to assess performance by analysing memory usage and the time required to identify frequent item groups. This research diverges from the norm by utilising four distinct datasets with varying sizes and transaction volumes, in contrast to the typical practise of using only two datasets for comparison. Following the independent evolution of algorithms, it was noted that the Apriori method exhibited longer execution time, while the Eclat algorithm required a larger amount of memory to identify frequent items across the four distinct datasets. The FP Growth algorithm demonstrated superior efficiency in terms of both time and memory consumption when compared to the Apriori and Eclat algorithms.

Initially, the study presents a highly efficient binary algorithm designed for extracting association rules from vast databases. The technique performs a single check on the initial database and utilises binary data representation and the vertical database layout concept during the mining process. Furthermore, the study conducted a comparison of the algorithms and demonstrated that FP-growth can serve as a viable option to address the performance limitations of Apriori, given its status as one of the most efficient mining techniques. FP-growth employs a concise and effective tree structure known as the Frequent Pattern Tree (FP-tree). Unlike the Apriori approach, which depends on the generation and testing of candidate item sets, FP-growth utilises a pattern growth technique to extract common item sets. Furthermore, their suggested methodology has the potential to reduce both execution time and primary memory utilisation in comparison to certain established algorithms [9].

The study utilises data mining tools and comparison algorithms to identify instances where diseases occur together, including correlation mining and the Apriori algorithm. Data mining is employed to analyse large volumes of data from many sources and generate valuable insights. Subsequently, this data is transformed into knowledge that can be analysed and understood. This substantial volume of data is utilised for the storage of healthcare information. Data extraction techniques are utilised to uncover suitable analyses and comparisons of patient data in this clinical dataset. Furthermore, it shown that employing an apriori approach will yield greater advantages when analysing medical data [11].

In their study, researchers [12] utilised the correlation rule-based Apriori algorithm to extract knowledge from medical repositories in order to forecast diseases. Additionally, they aimed to examine different data extraction approaches to identify connections between diseases. Now, let's examine three correlation mining algorithms together with their respective examples: AIS, SETM, and Apriori. The evaluation is grounded on several performance metrics, including support, accuracy, first stage speed, and computational speed. Prior work has posited that the Apriori method is beneficial for analysing medical data. Upon doing a thorough comparison, it was determined that the SETM algorithm outperforms all other algorithms that were examined.

Association rule mining is the most effective method of data mining for extracting information from large repositories and revealing concealed patterns contained within them, according to the study's findings [14].

This paper serves as a review of association rule mining (ARM), which is a descriptive technique in data mining. ARM is used to find and extract correlations, linkages, and interesting patterns in item sets. The research also provides a concise appraisal of the present state of frequent pattern mining. The frequent pattern mining technique is categorised into four distinct types: Association rule mining, itemised mining, sequential rule mining, and sequential pattern mining. The types of Association Rule Mining are Single-dimensional, Multidimensional, Quantitative, and Boolean association rules. Furthermore, it ultimately offers a comprehensive analysis of several algorithms by comparing their performance [13]. And this study primarily compared and analysed various association rule mining algorithms, including Apriori, FP Growth, Rapid Association Rule Mining (RARM), ECLAT, and Associated Sensor Pattern Mining of Data Streams (ASPMS) algorithms. The comparison focused on the algorithm's structure, database scanning time, and execution time. The objective was to gain insights into the distinctive features, strengths, and weaknesses of these algorithms by conducting the analysis on a consistent dataset. The primary challenges encountered by frequent pattern mining algorithms are database scanning and the development of intricate candidates. FP growth demonstrates a reduced number of database scans (2 scans), shorter execution time, and employs the divide and conquer technique [15].

The significance of comprehending the correlation between diagnosis and diagnostic tests is discussed in Paper [18]. This is particularly crucial in emergency rooms, where the urgency and expense of medical care are exceptionally high.

The emergency room's medical data was analysed using association rule mining to uncover concealed patterns and investigate the relationship between diagnostic testing requirements and diagnostic testing. The Apriori algorithm was employed as a mining technique for extracting these rules. After the completion of the therapies and the establishment of the rules, they underwent evaluation by specialists and practitioners in the emergency room. Recognising the correlation between the patient's diagnosis and the diagnostic criteria might assist emergency departments in making more informed judgements and optimising resource allocation. Additionally, these guidelines can be employed by doctors to effectively treat their patients. Furthermore, as indicated in Reference [19], the primary objective of this article was to demonstrate the methodology for managing substantial volumes of clinical data through the utilisation of data mining techniques. The significance of clinical data in the medical industry is paramount. Data mining is a valuable tool in the field of big data analysis, particularly for identifying patterns within extensive collections of general medical data. The primary focus of the paper involved elucidating data extraction techniques and contemplating their practical applications. This will have a direct impact on the comprehension of clinical researchers regarding the significance of data mining and the necessity to promote research advancement. This can have a significant impact on both clinicians and patients, resulting in increased satisfaction, reduced expenses and losses, and enhanced productivity.

In reference to the study mentioned in [16] An extensive analysis of the FP growth algorithm is conducted to identify and focus on the association rules among item sets. Alternative Association rule algorithms exhibit deficiencies and limitations in the processes of Database scanning and candidate key generation, resulting in significant time and memory consumption and heightened complexity when dealing with large datasets. In order to leverage the advantages of the F-P growth algorithm, this guide employs the 'Divide and Conquer' methodology, which eliminates the need for candidate critical generation tests and minimises the number of database scans (limited to only two scans). In addition, the apriori and F-P growth algorithms were compared across seven domains: storage architecture, search type, technology, number of database scans, memory utilisation, sparse or dense database, and runtime. The algorithms were implemented on two extensive and well used datasets: mushrooms and supermarkets. The analysis of the algorithms reveals that the f-p growth algorithm outperforms apriori in terms of both the time required to extract correlation rules and the memory usage for method implementation [17].

The study investigated Sequential Pattern Extraction (SPM), which distinguishes itself from current approaches through two key characteristics: continuity and the inherent unpredictability of path data. The SPM algorithm identifies sequences that have a higher occurrence rate than a minimal threshold provided by the user. The study chose specific algorithms to analyse and explore representational sequence patterns. These algorithms were then assessed using different factors, with the most significant ones being the time it takes to run the algorithm and the amount of RAM it consumes [20].

Among these strategies, the frequent pattern growth (FP growth) algorithm is the most efficient for discovering exciting rules. It works like this: F-P Growth processes the data by scanning the itemset just twice, giving the algorithm the advantage of increased time efficiency and memory savings. It does, however, have a slight flaw: It generates a large number of conditional FP trees. Finally, it presented a new, better FP tree that can decrease conditional FP trees, saving memory and time.

Data mining in the health sector is examined using research articles from the literature. To predict diagnoses based on symptoms, several research employ data mining techniques such as classification and clustering. Others, on the other hand, investigate association rule mining to generate rules that reveal intriguing correlations between components. Some studies compare the performance of various algorithms. Furthermore, other articles examine various data mining techniques, such as correlation mining and Apriori algorithms, and argue that the Apriori algorithm may be more useful for medical data. Based on the results of the survey, we determined the efficacy of various machine-learning algorithms in predicting cancer patients. As a result, in this article, we used frequent pattern analysis and association rules to find hidden patterns and associations in cancer data in Sudan that standard techniques could not reveal.

3 METHODOLOGY

The study's approach is founded on four distinct stages, as seen in Figure 1. This section provides a comprehensive overview of the methodology employed.

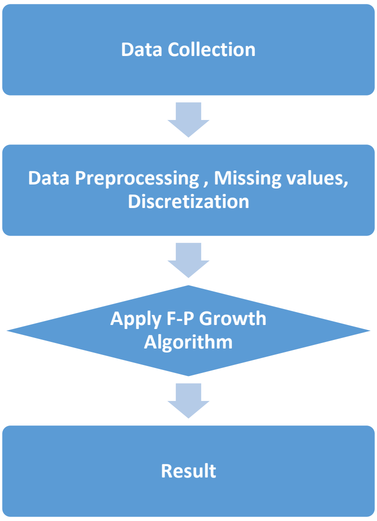
Data Collection
Data Preprocessing , Missing values, Discretization
Apply F-P Growth Algorithm
Result

Fig. 1. Data mining analysis process

3.1 Dataset

This study meticulously curated data from the statistics department of Khartoum Ontological Hospital. The database has 40,069 documents pertaining to cancer patients, spanning the period from 2015 to 2020. The extensive dataset comprising diverse cancer patients observed over an extended period enables a comprehensive examination of cancer trends and patterns. The dataset contains nominal and ratio properties, rendering it valuable for data mining. Features such as cancer kind help identify patterns and connections between different types of cancer. The duration of the illness and the individual's age can serve as indicators of the severity and progression of the condition. The analysis encompasses the most recent facts from 2015 to 2020, rendering it significant. This guarantees that the study's results accurately represent the current cancer prevalence and incidence rates in Sudan

. Table (1) displays a sample of the dataset, while Table (2) exhibits the dataset after age discretization.

Table (1): Sample of Dataset

Serial No	Year	Gender	Age	State	Diagnosis
14875	2015	Female	40	South Darfour	NPH
112130	2015	Male	4	Khartoum	Willims tumour
112132	2015	Female	1	South Darfour	Retinoblastoma
112133	2015	Male	4	South Kurdufan	CML
112134	2015	Male	70	El- Jazeera	Gingiva

3.2 Data Preprocessing

Real-world data frequently exhibits issues such as incompleteness, noise, or inconsistency. Examining such data yields deceptive outcomes. Data pre-processing is an essential and pivotal stage in the data mining process.

3.2.1 Missing values

The dataset has some values in some attributes that need to be added. This can hurt the generated rules. The Replace Attribute operator substitutes these absent values by initially utilizing the Filter operator, with the condition class parameter set to 'no missing values'.

3.2.2 Discretization

The attribute "age" contains a wide range of values, ranging from 1 to 117 years. To simplify the data, it is necessary to discretize the attribute by dividing it into a series of consecutive intervals, or "bins," each representing ten years. The bin width can be calculated using the formula: Bin width = (Max–Min) / Square Root of the Data points.

Table (2): Data set after age discretization

Serial No	Year	Gender	Age	State	Diagnosis
14875	2015	Female	30-40	South Darfour	NPH
112130	2015	Male	0 – 10	Khartoum	Willims tumor
112132	2015	Female	0 – 10	South Darfour	Retinoblastoma
112133	2015	Male	0 – 10	South Kurdufan	CML

3.3 F-P growth

FP-growth is a method for increasing patterns that utilizes a specialized data structure called FP-tree. FP-growth algorithm identifies frequent item sets by identifying all frequent items in 1-itemsets within the context pattern base. The situation pattern foundation is effectively produced by connecting the node shape and the FP-tree. FP-boom does not explicitly produce candidate item sets [21].

The FP growth process uses transaction data to construct frequent item sets, creating an FP tree. Each tree node represents a dataset item and its corresponding frequency count. The tree originates from the root node, devoid of anything. The elements in the dataset that often appear together with a node in the FP-tree are referred to as its progeny.

The fundamental concept of the FP growth algorithm can be described as an iterative elimination process: during a pre-processing phase, each object is individually eliminated from infrequent parameters.

There are two distinct stages:

Step 1: Construct a condensed data structure called the FP-tree (Frequent pattern tree).

Step 2: Involves extracting frequent item sets directly from the FP-tree by traversing across it.

The benefits of FP-Growth are as follows:

- The dataset is processed in only two iterations.
- Condenses dataset.
- Elimination of candidate generation.
- Significantly more efficient than Apriori.

3.4 Measures

The paper uses two measurement metrics: support and confidence. Support is used to estimate the occurrence of an item or set of items among the total number of transactions. In contrast, confidence is used to measure the strength of the relationship between groups of items.

3.4.1 Support

Support is a measurement parameter utilized to identify the occurrence of an item or a set of items within a given set of transactions. It can be defined as the frequency with which an item or set of items appears in the transactions. An item or set of items is considered frequent or extensive if its support value is higher [22]. Utilizing the likelihood notion, we can define support as:

$$\text{Support} = \text{P}\,(\text{A} \cap \text{B}) = \frac{\text{number of transactions containing both A an B}}{\text{Total number of transactions}} \tag{1}$$

Where A and B represents the itemsets in a database D.

3.4.2 Confidence

(2)

3.4.3 Confidence is a measure used to evaluate the strength and association between groups of items. It assesses the likelihood of item B occurring in the same transaction as item A. In simpler terms, confidence helps determine the conditional probability of item usage [19]. Utilizing the likelihood notion, we can define confidence as:

$$\text{Confidence} = \text{P}\,(\text{A}\,|\,\text{B}) = \frac{\text{P(A} \cap \text{B)}}{\text{P(A)}} = \frac{\text{number of transactions containing both A an B}}{\text{number of transactions containing A}}$$

4 RESULTS AND DISUCSSION

The rules were derived from the FP growth algorithm model, with a minimum support of 0.05 and a minimum confidence of 0.5. The model efficiently produces 70 rules of exceptional quality within a short timeframe with minimal memory usage. Additionally, it develops rules that exhibit the highest levels of confidence and correlation across multiple factors.

These rules are the most ideal ones generated considering criteria *(Age, State, Diagnosis → Gender)*

1.	[State = Khartoum, Diagnosis = Prostate]	-->[Male]	(confidence: 0.994)
2.	[Diagnosis = Cervix]	-->[Female]	(confidence: 0.993)
3.	[State = Khartoum]	-->[Female]	(confidence: 0.575)
4.	[Age = 40-50, State = El-Jazeera]	-->[Female]	(confidence: 0.630)
5.	[State = Khartoum, Age = 50-60]	-->[Female]	(confidence: 0.632)
6.	[Diagnosis = Ovary]	-->[Female]	(confidence: 0.995)
7.	[Age = 70-80, Diagnosis = Prostate]	-->[Male]	(confidence: 1.000)
8.	[Age = 60-70, Diagnosis = Prostate]	-->[Male]	(confidence: 0.994)
9.	[Diagnosis = Prostate]	-->[Male]	(confidence: 0.995)
10.	[State = Khartoum, Age = 70-80]	-->[Male]	(confidence: 0.587)
11.	[Age = 40-50, Diagnosis = Breast]	-->[Female]	(confidence: 0.973)
12.	[Diagnosis = Breast, Age = 30-40]	-->[Female]	(confidence: 0.986)
13.	[State = Khartoum, Age = 50-60, Diagnosis = Breast]	-->[Female]	(confidence: 0.967)
14.	[State = Khartoum, Age = 40-50, Diagnosis = Breast]	-->[Female]	(confidence: 0.979)
15.	[Diagnosis = Oesophagus]	-->[Female]	(confidence: 0.552)
16.	[Diagnosis = CML]	-->[Male]	(confidence: 0.524)
17.	[Diagnosis = Lymphoma]	-->[Male]	(confidence: 0.606)

The results clearly and indisputably show that the FP growth technique is superior in terms of implementation speed and memory utilisation. By thoroughly analysing past trials, we utilised this method to extract significant observations from the data, which served as the foundation for our conclusions.

Figure 3 depicts the distribution of cancer cases across the several states of Sudan, while **Figure 4** specifically highlights the prevalence of infection among males, with the highest rate of 27.8% observed in Khartoum state. Lymphoma comprises 4.3% of cases, whereas chronic myelogenous leukaemia (CML) accounts for 3.7%. **Figure 5** illustrates the five most common kinds of cancer in Sudan. The most prevalent malignancies in both males and females are breast, prostate, ovary, oesophagus, and cervical cancers. Out of all the types of cancer, breast cancer is the most common among women, accounting for 29.4% of all occurrences. Ovarian cancer has a prevalence rate of 8.1%, followed by cervical cancer at 6.8%, and oesophageal cancer at 4.4%. **Figure 2** illustrates the gender-based distribution of cancer cases, with women representing 55.1% and men representing 44.9% of the overall instances. **Figure 8** offers valuable insights into the geographic spread of cancer cases in different regions of Sudan, classified by age and gender. The prevalence of breast cancer is highest in Khartoum state, with a rate of 31.1%, followed by Al-Jazeera state with a rate of 8.7%. **Figures 6 and 7** display the occurrence of cancer in various age categories, together with rates categorised by age group and gender. The highest prevalence of cases is observed among women between the ages of 40 and 70, representing roughly 66.9% of the total. Prostate cancer is the predominant form of cancer in men, accounting for 13.3% of all cases. The condition mostly impacts adults aged 60 to 90, with a prevalence rate of 80.7% within this age range.

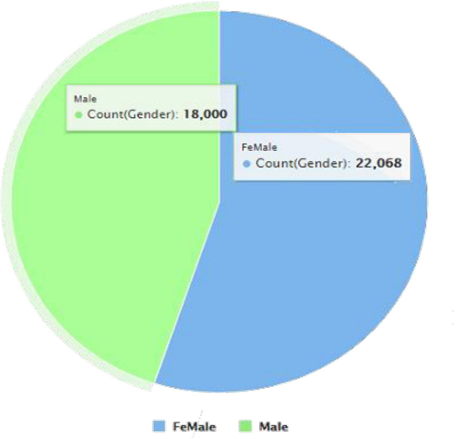

Fig.2. Distribution of cancer with Gender

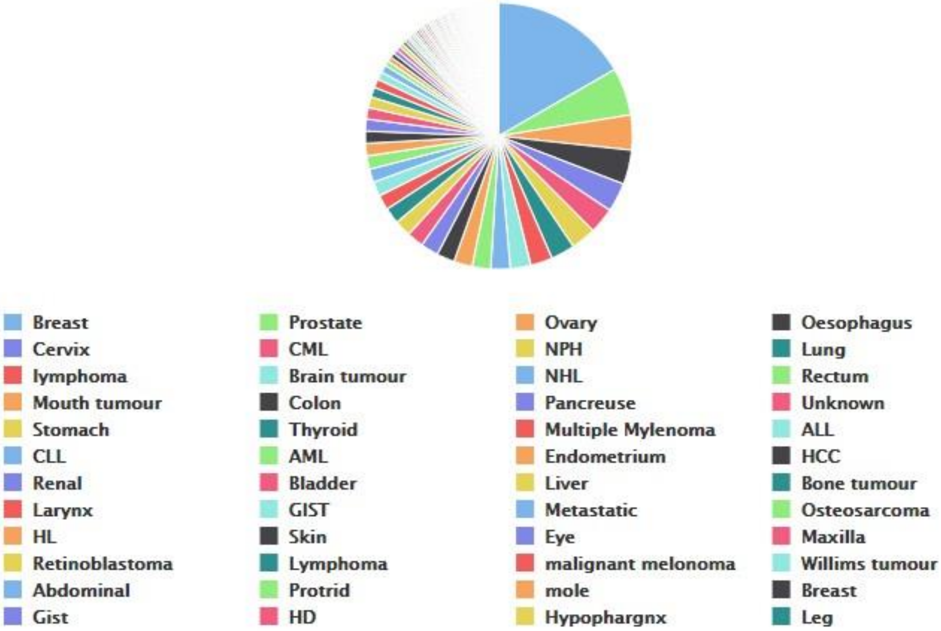

Fig. 3. Distribution of cancer incidences within Sudan States

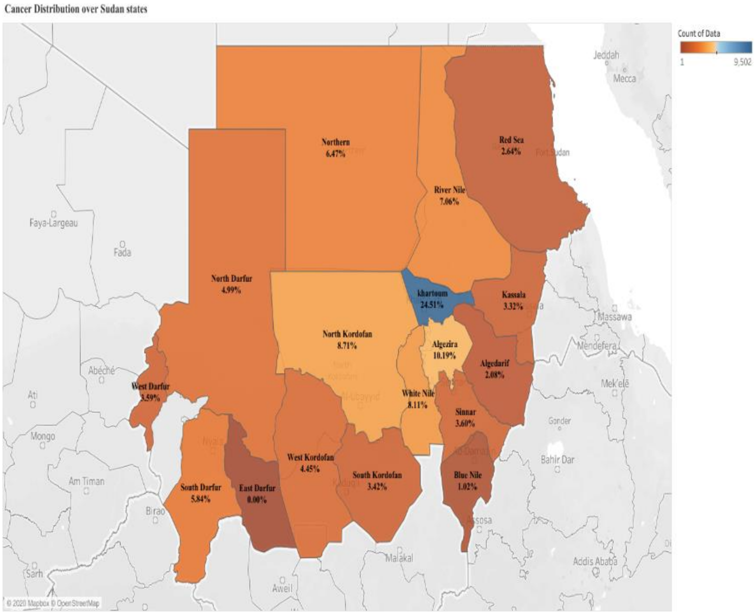

Fig. 4. Distribution of cancer incidences within Sudan States

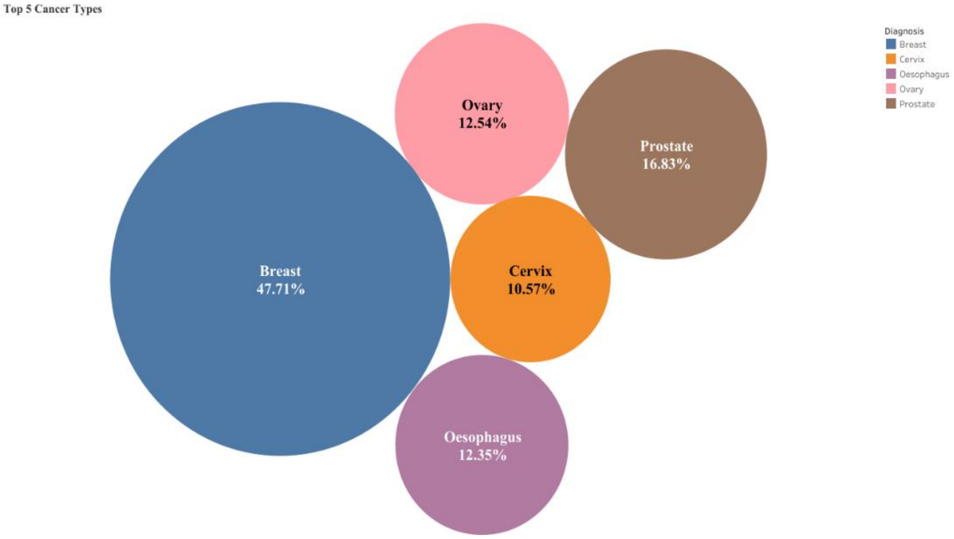

Fig. 5. Top 5 cancer incidences within Sudan States

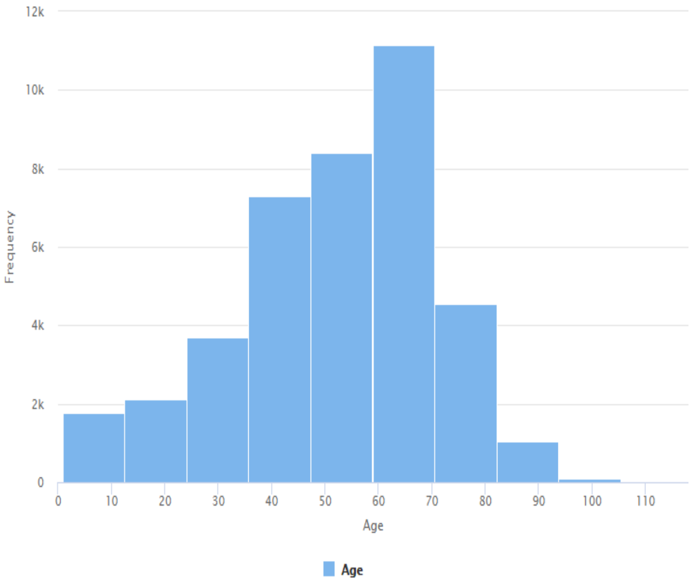

Fig. 6. Distribution of cancer within age groups

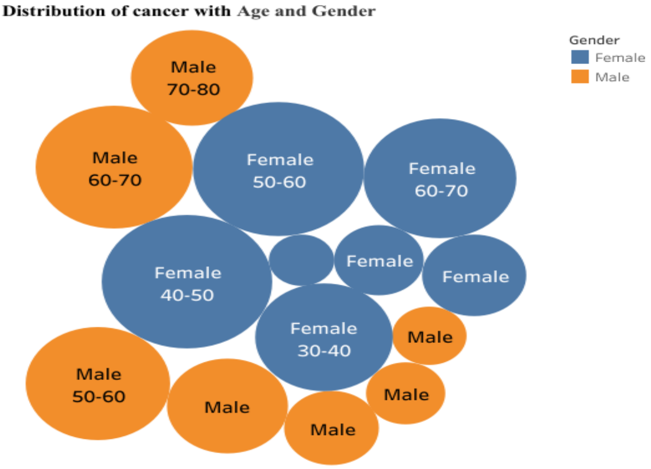

Fig. 7. Distribution of cancer rates based on group of Age vs. gender

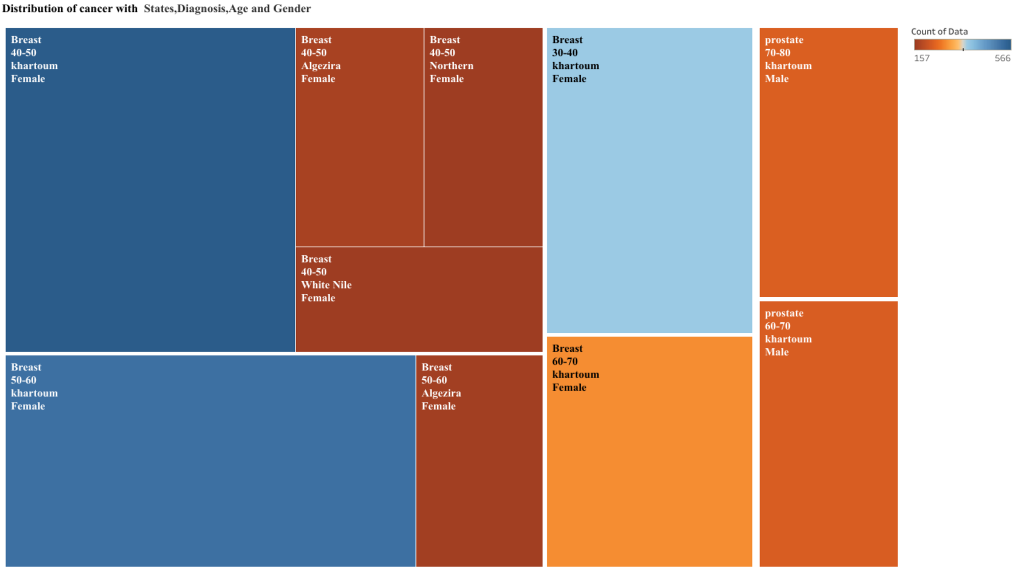

Fig.8. Distribution of cancer occurrences in Sudan States vs. age and gender

5 CONCLUSION

This paper explores cancer in Sudan using data mining approaches, focusing on the FP growth algorithm. An analysis of a large dataset of cancer cases from the Sudan Cancer Registry reveals essential trends and correlations that can give policymakers and healthcare providers crucial information. The study discloses that the rate of cancer is more frequent in women, with breast cancer being the most extensive. In addition, Khartoum state has the highest rate of cancer cases among both males and females in terms of geographical distribution. The age group most impacted for women is between 40 and 70 years, whereas for men, it is between 60 and 90. The FP growth algorithm has superior efficacy and scalability compared to the Apriori algorithm, making it the optimal tool for obtaining valuable insights from data. The results of this study show that the FP growth algorithm can extract information from vast and intricate datasets in the healthcare field. These findings serve as a helpful reference point for future research and practical use. The report suggests that data mining techniques should be implemented in the Sudanese medical sector to extract valuable insights from the extensive information stored in the Sudan Cancer Registry. Sudan should implement a comprehensive and uniform electronic health record system for patients, enabling efficient data analysis and exploration and enhancing clinical decision-making and personalized care. The study proposes the establishment of a sophisticated computational infrastructure in Sudan to facilitate the analysis and extraction of information from extensive and intricate databases. Further investigation is warranted to examine the application of hybrid and meta-heuristic algorithms to enhance the optimization of parameters and performance of the FP growth algorithm and other data mining approaches. The discoveries presented in this research enhance the comprehension and control of cancer in Sudan, ultimately resulting in improved health outcomes for those diagnosed with cancer. The research concludes by emphasizing the importance of data mining in healthcare and its potential to transform patient care and outcomes.

Acknowledgement:

The authors extend their appreciation to the University Malaysia Sabah for funding this work through SKIM PENSYARAH LANTIKAN BARU (Project under grant number (SLB2256)).

REFERENCES

[1] *Fran Boyle AM, "Introduction to Cancer", Cancer Council Australia 2017.*

[2] *Marc B. Garnick, MD , Beth Israel Deaconess Medical Center "An Introduction to Cancer and Basic Cancer Vocabulary ", Harvard Medical School, Boston.*

[3] *K. J. Cios, "Data Mining: A Knowledge Discovery Approach", Springer, 2007.*

[4] *J. Han, Micheline Kamber and Jian Pei, "Data Mining Concepts and Techniques", Waltham: Morgan Kaufmann 2012.*

[5] *Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, Ahmedin Jemal, " Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries".IARC 2018.*

[6] *Lakshmi K.Sa , G.Vadivu ,"Extracting Association Rules from Medical Health Records Using Multi-Criteria Decision Analysis", 2017.*

[7] *Ramah Sivakumar, J.G.R.Sathiaseelan,"A Performance based Empirical Study of the Frequent Itemset Mining Algorithms",ICPCSI-2017.*

[8] *Sallam Osman Fageeri, Rohiza Ahmad, Baharum B. Bahraini,"BBT: n efficient association rules mining algorithm using binary-based technique",2014.*

[9] *Marwa Maweya Abdelbagi ElbasheerID, Ayah Galal Abdelrahman AlkhidirID, Siham Mohammed Awad Mohammed, Areej Abuelgasim Hassan Abbas, Aisha Osman Mohamed, Isra Mahgoub Bereir, Hiba Reyad Abdalazeez, Mounkaila Noma"Spatial distribution of breast cancer in Sudan 2010-2016" Alzaiem Alazhari University, Khartoum, Sudan, 2016.*

[10] *J. Sabthami, K. Thirumoorthy and K. Muneeswaran, "Mining Association Rules for Early Diagnosis of Diseases from Electronic Health Records",2016.*

[11] *D. Sheila Freeda, and M. Lilly Florence," An 0verview of Disease Analysis using Association Rule Mining ',2017.*

[12] *[D. Sheila Freeda, and M. Lilly Florence,"An 0verview of Disease Analysis using Association Rule Mining ",2017.*

[13] *Thabet Slimani, Amor Lazzez," Efficient Analysis of Pattern and Association Rule Mining Approaches", Taif University, KSA.*

[14] *Meera Narvekara, Shafaque Fatma Syedb, An "optimized algorithm for association rule mining using FP ",2015.*

[15] *Shamila Nasreen, Muhammad Awais Azamb, Khurram Shehzada, Usman Naeemc, Mustansar Ali Ghazanfara," Frequent Pattern Mining Algorithms for Finding Associated Frequent Patterns for Data Streams: A Survey " ,2014.*

[16] *Shivam Sidhu D, Upendra Kumar,Aditya Nawani ,FP Growth Algorithm Implementation ,India ,2014.*

[17] *M.S. Mythili,A.R. Mohamed Shanavas, "Performance Evaluation of Apriori and FP-Growth Algorithms" ,2013.*

[18] *Görkem Sarıyer, "Highlighting the rules between diagnosis types and laboratory diagnostic tests for patients of an emergency department: Use of association rule mining " ,2019.*

[19] *Wen-Tao Wu, Yuan-Jie Li, Ao-Zi Feng, Li Li, Tao Huang, An-Ding Xu , Jun Lyu "Data mining in clinical big data: the frequently used databases, steps, and methodological models",2021.*

[20] *Shiting Ding , Zhiheng Li , Kai Zhang , Feng Mao "A Comparative Study of Frequent Pattern Mining with Trajectory Data",2022.*

[21] *Mohammed Al-Maolegi, Bassam Arkok "An improved apriori algorithm for association rules",2014.*

[22] *J. Han, Micheline Kamber and Jian Pei, Data Mining Concepts and Techniques, Waltham: Morgan Kaufmann 2012.*