ISSN 2622-626X

On the adaption of data mining technology to categorize cancer diseases

Manal Al-Dafas [a,1,*], Ammar Albujeer [b,2], Shaymaa Abed Hussien[c,3] , Raed Khalid Ibrahim[d,4]

[a] Information Science Departement, Faculty of Arts and Humanities, King Abdulaziz University, Jeddah, Saudi Arabia
[b] Nab'a Al Hayat Foundation for Medical Sciences and Health Care, Najaf, 54001, Iraq
[c] Al-Manara College for Medical Sciences, Maysan, 62001, Iraq
[d] Department of Medical Instruments, Medical Technical College, Al-Farahidi University, Baghdad, 10022, Iraq.
[1] maldafas0001@stu.kau.edu.sa*; [2] ammar.dent@yahoo.com; [3] shaimaa2021@uomanara.edu.iq; [4] raad.khalid@alfarahidiuc.edu.iq
* corresponding author

ARTICLE INFO

Article history
Received: 2021-12-05
Revised: 2022-02-20
Accepted: 2022-06-29
Published: 2022-12-20

Keywords
Data Mining
Disease Categorization
Categorization Algorithm
Cancer Diseases
Prediction

ABSTRACT

Along with data mining, tools and software have emerged to aid in mining the vast and growing amount of data to access knowledge in databases. These tools facilitate work on most scientific disciplines, including sciences, Libraries and information. Accordingly, Data mining became an effective technique for obtaining knowledge to achieve the basic goal of discovering hidden facts that are contained in databases through the use of multiple technologies that include artificial intelligence, statistical analyzes, techniques and data modeling etc. Medical data mining is considered one of the most important tools used in the field of medicine, especially in exploring and knowing health conditions according to records of former patients. In addition, data mining helps not only in categorizing cancer but also in taking the necessary measures. With the spread of cancer at high rates around the world, the need to develop smart methods that have the ability to predict the disease appeared. Applications of data mining techniques spread as human attempts to control this deadly disease, with the aim of awareness, early detection and reduction of treatment costs. This prompted the researcher's curiosity to know the ability of data mining to categorize cancer. This work aims at reviewing ways to solve one of the problems that doctors suffer from, which is the problem of diagnosing diseases that lead to death, including cancer, as there is huge information that has not been used. Therefore, this work tries to solve this problem using data mining technology in addition to helping doctors make the right decision. The study reached several conclusions, namely the fact that the studies presented in the paper demonstrated the effective role of data mining techniques in reducing medical errors in terms of their ability to predict and accurately diagnose the disease, as well as the effectiveness of the algorithms of the data mining technique in predicting the presence of the disease at an early stage. Thus, we found that the clinical field needs to expand research, foster new kinds of calculations, and apply them practically speaking to create the best and most precise outcomes and even to supplant or surpass specialists' performance at this level..

1. Introduction

Major developments in information technology have led to the excessive growth in health informatics data as health computing includes hospital details, patient details, disease details and

treatment costs. This big data is obtained from various sources and in various formats. These can be disorganized features and missing data. Applying data mining techniques is a major approach to extract knowledge from major diseases and from their symptoms. Data research contains various methods of extracting knowledge from a wide range of disease-related data. Data mining techniques such as classification, grouping, and rules can be used to analyze data and extract useful information.

Data mining is the discovery of information from Data mining or data (sometimes called knowledge discovery) is the process of data analysis from different perspectives, extracting relationships between them, and summarizing them into useful information, such information can contribute to increasing profit, reducing costs, or both. Or It is the process of uncovering and finding information of interest through the use of a group It is a complex tool. Some of these tools include regular statistics tools Artificial intelligence and computer-made infographics.

Some of the important applications of research in health include predicting future diseases based on past data collected in similar diseases, diagnosing diseases based on patient data, analyzing treatment costs and resource applications, lost data and reducing disease waiting time diagnose. Data search tools such as Weka, Fast Miner and Orange are used to predict and extract new data, and the problem of the study lies in the spread of many types of cancer and the inability to predict and identify the disease, which would improve the cost-effectiveness of health services and reduce the time required to diagnose the diseases [1]. Thousands of people die every year due to delays in diagnosis and detection; therefore, this study comes to reveal the tools that help improve the correct diagnosis of diseases and reduce time and pressure on doctors, which is the technique of data mining as there is a huge data that has not been utilized. Thus, this study came to demonstrate the impact of this technology on the ability of cancer early detection. From this standpoint, the idea of this article emerged, which has the main question: What is the effect of using data mining technology to categorize cancer diseases by using the method of critical evaluation through literature and related studies review????

The importance of this work lies in the effectiveness of data mining and data extraction and its role in early detection of cancer and reducing treatment costs. Scientifically, technology has a tremendous ability of to face some of the biggest challenges in analyzing huge information through patients' health data. Moreover, data mining plays a leading role in reducing medical errors in diagnosing the disease. In practice, the importance of this work appears through the ability of data mining to help doctors make the right decision at the right time.

The number of data mining programs in the market is growing exponentially, so there is need to choose a standard for a software package that can be made available to intended beneficiaries and organizations. With the continued increase in the number of programs and the additional benefits they contain Newer programs, it becomes more difficult to choose the right software package, as it may be connected Wrong decision with losing a lot of time and money, so many studies have been conducted around the world to evaluate programs; However, the researchers did not reach the beneficiaries to generalize the standards Selection and evaluation. Improper selection of the software package can be extremely costly. It negatively affects the work.

2. Related Work

There are hundreds of studies on data mining, but they focus on Mostly business and statistics, and there are a few studies that provide. It deals with the evaluation of data mining software and tools and how to select them, except here general studies. The main focus of this section aims to discuss and analyze some previous studies related to the subject of the study in a critical and analytical way. In [2] "Applying Data Mining to Investigate Cancer Risk in Patients with Pyogenic Liver Abscess" (2020), this review targets setting up preventive intercessions by creating rules for the avoidance and early location of liver malignant growth. It uncovers the capacity of information mining to arrange liver disease. The main role of this review is to dissect the dangers of CRC and HCC in patients with PLA, the variables that add to malignant growth advancement for these patients, and clinical tests to analyze these tumors. Clinical records and clinical information were recovered from the data set. Insights and information mining innovation have likewise been utilized to distinguish potential danger factors related with malignant growth illnesses. Ideal models for choice tree examination have been distinguished to assist clinicians with deciding proper clinical tests for early conclusion of these diseases and furthermore for an expanded solid life expectancy.

The patient information was acquired from the Clinical Informatics Research and Development Center, (Taichung) Taiwan. The information securing period was from January 1, 2006, to December 31, 2013. The exploration test was separated into two gatherings: trial and control gatherings. The exploratory gathering comprised of tainted in patients with an underlying finding. Patient information, including age and sex, were followed from the date of the primary finding of liver boil and followed for a very long time. The benchmark group was chosen based on blending with the trial bunch as per age and sex. The extent of members from the control test bunch was 1:10. All patients in the emergency clinic from 2006 to 2013 were tested arbitrarily. Then, at that point, they followed their clinical information from one year preceding the date of starting analysis and followed them up from there on until 2013.

All factual examines were performed utilizing SPSS as indicated by the measurements of the Ministry of Health and Welfare, the rate of malignant growth is influenced by age and sexual orientation. In this manner, chances proportions (ORs) for malignancy in the test bunch (with liver sore) ought to be determined by age and sexual orientation. Contrasts between bunches in the impacts old enough and still up in the air utilizing t-tests and chi-square, separately. In the wake of ensuring that there were no genuinely critical contrasts in age and sex between the trial and control gatherings, and the occurrence rate, CRC and HCC were investigated. Chances proportions were determined utilizing calculated relapse, CRC and HCC hazard investigation. The review brought about that the chances proportions for malignancy in the trial bunch were higher than those in the benchmark group. Accordingly, the dangers of CRC and liver disease in PLA patients were higher than that of the ordinary individual. Accordingly, PLA was utilized as an exploration variable in the prescient model, which was contrasted with the forecast model utilizing PLA as a grounded expectation model. they utilized the C5.0 choice tree technique in SPSS to extricate grouping rules from the information.

The consequence of this review shows that PLA patients had higher paces of HCC and CRC than other hospitalized patients. Consequently, generally colorectal and liver diseases are distinguished in the center and late stages. Be that as it may, when tumors are trapped in the beginning phases, endurance rates will be high. The characterization rules for CRC can be classified by less factors and have a more precise anticipation for this. The potential danger factor for liver cancer is key in distinguishing people in danger of creating liver malignant growth. The AFP tumor index is used primarily to test for liver cancer. Therefore, when caring for patients with PLA, physicians must be careful with hemoglobin. Finally, there are no studies confirming that there is an association between PLA and CRC or HCC. However, this relationship requires further exploration. There are certain risk factors for cancer, such as living habits, genetics, and alcohol use that cannot be identified in the clinical database.

Regarding article [3] entitled "Make Intelligent of Gastric Cancer Diagnosis Error in Qazvin's Medical Centers: Using Data Mining Method" (2019), we found that this study aims at assuring the possibility of exploring data, techniques, and characteristics of disease risk factors for predicting and diagnosing stomach cancer. In fact, stomach cancer is one of the most common types of cancer, as it is considered a high risk if misdiagnosed or diagnosed late. This applied research was carried out by a retrospective analytical descriptive method among individuals who were referred to the selected health centers in Qazvin city in 2017. In this paper, the proposed model is based on CRISP, which consists of six stages. Each of these stages consists of subsections. These six phases include Business, Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Dissemination. A number of (405) participants were selected from the two groups of healthy individuals and patients transferred to Boalie Hospitals. The records of the participants contained complete data for analysis in the databases that included the participants' gender, age, weight loss, abdominal pain, nausea, anorexia, dysphagia, pernicious anemia, and Milena. The data collection tool was a data extraction model that was designed based on the characteristics used in the assay to solve the problem of class imbalance. The results of this study were also obtained on the basis of a 10-fold validation technique. Accordingly, the entire database is initially divided into training and test sets, then the training set is divided into 10 parts. In each iteration of the cross-validation process, one part was identified as the validation group, while the rest of the data was selected as the training group. We note that 70% is being defined from the percentage of total database samples as the training set and the remaining 30% as the test set. In this study, learning methods were used because in a medical application, a very large number of factors (characteristics) usually accompanies the medical information of a patient and

taking into account all these characteristics by the physician makes it difficult to make a decision about the patient's condition.

In addition, when applying mathematical methods, error and complexity appear, and thus leading to low efficiency. In view of the large number of characteristics in stomach cancer patients, this issue is of particular importance. For this reason, the four learning methods of SVM, DT, Bayesian Model and KNN were studied to categorize patients with stomach cancer. The aim of this study is to determine which of these four methods of machine learning has the highest accuracy in categorizing stomach cancer samples. The results of the study, which were verified to evaluate the four methods of the SVM, DT, Bayesian model and KNN algorithms were 90.08, 87.89, 87.60 and 87.60 percent, correspondingly. Likewise, the outcomes showed that the most elevated F score was related with SVM (91.99); while the least rate was related with the KNN calculation (87.17). In this way, the aftereffects of the review were that the SVM calculation is the most incredible in arranging test tests. Subsequently, this shrewd framework can be utilized as a clinical right hand in the clinical field.

With regards to study [4] entitled "Application of data mining methods to improve screening for the risk of early gastric cancer" (2018), it aims at detecting stomach cancer disease through data mining. Consequently, the motivation behind this review was to fabricate prescient models to look at EGC chances dependent on a few elements, like segment attributes, dietary patterns, primary indications inside roughly 3 months, family ancestry or past illnesses and mental diagnostics of patients experiencing irresistible sicknesses. It also analyzes simultaneously the major effects on EGC risks, so as to help in making clinical decisions to raise the diagnosis level of the risk of infection with EGC even more.

In spite of the fact that stomach malignant growth is a metastasis and with high paces of patients and passings in China, the endurance pace of patients with early gastric disease (EGC) is high after careful resection. Besides, to upgrade the early finding and assessment of patients with EGC this review applied information mining techniques to further develop EGC hazard identification.

The substance of this review came from an undertaking called - "An inventive stage for early recognition of stomach malignancy dependent on distributed computing" in the primary clinic associated with Guangdong Pharmaceutical University from January 2016 to May 2017. The review test was 620 patients. A survey was utilized that included nine segment attributes, 11 dietary patterns, 14 significant side effects during around 90 days. Endoscopy was recorded notwithstanding obsessive biopsy, and the back line is the highest quality level in EGC test. Various (618) members were remembered for the first dataset. These members were arranged into generally safe of EGC (487 cases) and high danger for EGC (131 cases) as per their outcomes. Separated irregular examining dependent on the danger of EGC prepared the datasets. The first informational index was partitioned into a 70% preparing set and a 30% test set. The preparation set was utilized to make a model and a test set to assess the model.

At last, we would presumably get a decent sign of the degree to which the model would be summed up to other datasets that were like the current dataset since the low and high-hazard proportion of EGC was an unequal preparing set (patients at low and high danger of EGC were 344 cases and 98 cases individually). A lopsided order would diminish the expectation for each type of the classifiers, so the current review utilized the manufactured minority oversampling strategy (SMOTE) to adjust the preparation set. Destroyed contrasts from straightforward over-replacement and absence of tests which some past research showed. Destroyed successfully sped up the goal of classifiers, for example, support vector machine, C4.5 choice tree, Dom Forest runs, Bayesian organization and neural organization. Subsequent to managing the uneven order with SMOTE, preparing bunch tests expanded to 516 cases, with 344 cases at generally safe of EGC and 172 cases at high danger of EGC. The repetitive preparing set was utilized to create the expectation model.

This review utilized polls and test information to execute four models of EGC hazard screening. Three information extraction models with better execution could be applied to help clinicians in progressive determination of EGC hazard, which would further develop EGC screening at scale. Information mining models may rapidly survey the movement of stomach malignant growth, which will draw in the consideration of clinicians and patients, then, at that point some proper measures will be taken to improve endurance, particularly when patients are in danger of creating EGC. One of the main consequences of the review was that it discovered 16 significant impacting factors for the danger of creating EGC, like sort of occupation, HP contamination, HP antibodies, drinking boiling water,

eating cured food sources, and so on They are a token of early anticipation, early recognition and early therapy of stomach malignancy. Generally, this review might help clinical analysts in choosing and carrying out ideal prescient models and surveying significant effect factors.

In [5] the study entitled "Analysis of the factors influencing lung cancer hospitalization expenses using data mining" (2015), This study aims at classifying and analyze lung cancer hospital expenses to predict reasonable medical costs as it has become a common issue for both hospitals and insurance institutions as lung cancer treatment expenses are not only an economic burden on patients but also a burden on medical insurance companies. In this paper, data mining technology is widely used in the medical field, such as disease diagnosis, knowledge discovery of Chinese traditional medicine, treatment effect analysis, death probability prediction, survival prediction, improvement of medical quality, and medical insurance. The main objective of the paper is to discover the factors that affect hospital expenses for lung cancer patients and the role of different variables in expenses using data. The paper also formulates classification rules for medical expenses to provide a theoretical basis for monitoring hospital expenses and formulating its policies. Data mining refers to the process of discovering hidden and/or unknown, but potentially useful, information and knowledge from excessive, incomplete, and vague random data. This research is based on the C5.0 algorithm and uses the IBM SPSS MOLDER 14.2 program to analyze the factors affecting hospitalization expenditures from lung cancer patients.

The hypothetical premise of the C5.0 calculation is hypothetical data. The data obtaining proportion is utilized as rules to decide the best variations of grouping and dividing focuses. The calculation accepts the yield variable as the data U, which transmits from the data source and the info factors as a progression of V data that the data source gets. Before the choice tree is produced, the yield variable is totally arbitrary for the data source.

The utilization of the C5.0 calculation to decide the affecting variables and grouping rules for the expenses of treating patients with cellular breakdown in the lungs resolves the issues of expecting treatment expenses and confirming clinic consumptions. It is likewise a reference for patients in their decision of therapy and gives choice help to clinics to lessen the expense of clinical treatment administrations.

Also, it has given a clinical reference to the protection offices to decide the extent of medical clinic costs for cellular breakdown in the lung's patients. One of the consequences of this paper was to explain the impacts of various factors for the expense of treating patients with cellular breakdown in the lungs. The grouping rules for malignancy are in accordance with the clinical act of cellular breakdown in the lungs. Order rules made in the structure reveal emergency clinic costs for cellular breakdown in the lungs patients and might be applied to different infections.

Regarding [6] A study entitled "Preprocessing Breast Cancer Data to Improve the Data Quality, Diagnosis Procedure, and Medical Care Services" (2020), this review targets discovering the impact of pre-handling to give great information to classification methods to work on the presentation of classifiers dependent on the chose qualities. This review came to recognize bosom malignant growth through information mining. Accordingly, the outcomes showed that the proposed property choice technique was superior to the series drifting inquiry strategy and it prompted better execution of the classifiers.

In such article, a strategy has been proposed to foresee bosom and colon tumors in an uneven dataset. To begin with, to adjust the informational collection, the examples of the minority bunch were expanded through over-examining, and afterward, expectation models were utilized. In this review, the checking utilized a wide scope of strategies, henceforth, a bunch of fitting pre-treatment techniques were chosen thinking about the idea of the information determined to further develop order by shape. Hence, this review can possibly fill in as an aide for the use of fitting pre-treatment.

In this review, some pretreatment strategies, like mistake and remedy, settled information irregularity, invalid qualities were applied with the choice of RROC bosom disease dataset. Toward the start, just 40% of the RROC dataset quality qualities were filled in. The review meant to analyze the impact of pre-handling on further developing information quality just as the calculations' order results. The RROC dataset was gathered cross-sectionally from 2009 to 2014 utilizing patient records of bosom disease danger from RROC in Iran, and was physically gone into the SPSS program.

Additionally, in this review three characterization calculations were utilized which are Bayes, k-NN and SMO classifier. Weka application was utilized since it was a generally utilized application for these calculations. As referenced in Kerdegari et al, Weka is a Java-based information mining program. Weka contains a bunch of learning calculations for information mining undertakings like pre-handling, order, and property choice. The graphical interface of Weka is called Explorer, and all joined utilities can be gotten to through this UI.

Classifier Naïve Bayes. This algorithm is taken from Bayes. Forecast models are made by this calculation. The contextual analysis in this review was separated into 3 sections. Expectation models were produced in the initial segment utilizing the natural dataset, and in the second and third parts, forecast models were created utilizing the pre-arranged dataset.

2.1 Case study of data pre-processing at RROC

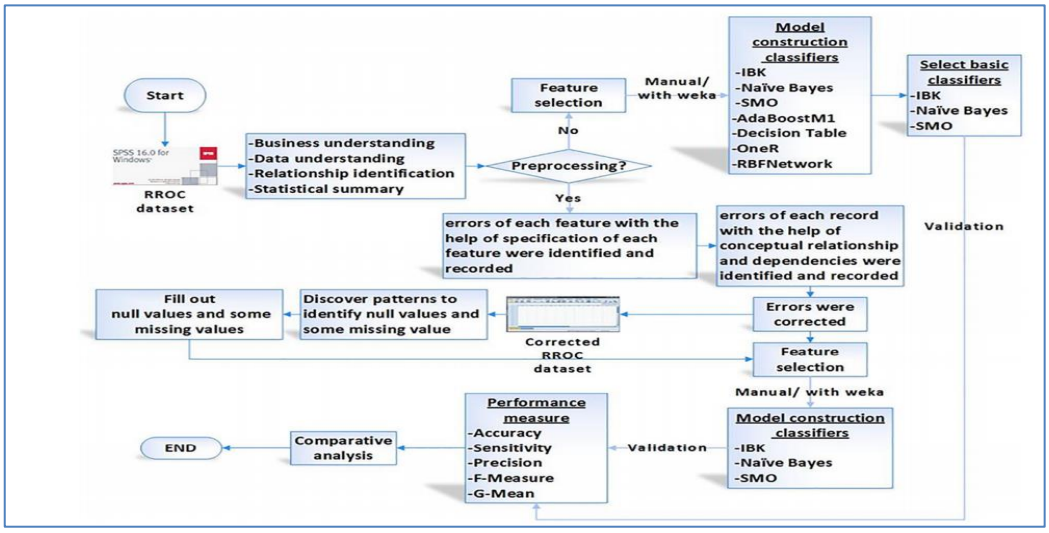

Fig. 1. The Steps of Case Study in RRCC

For making expectation models Weka 3.6.9 was utilized. In all executions, every exercise manual was run multiple times and afterward scores were arrived at the midpoint of to forestall inclination in a particular piece of information. During the time spent information extraction, the dataset is isolated into two pieces of the preparation set and the test set. The preparation set is utilized to gain proficiency with the calculations and the test set is utilized to test the order calculations. The experimental group expects that the while class esteems don't exist until the hour of the information mining assessment.

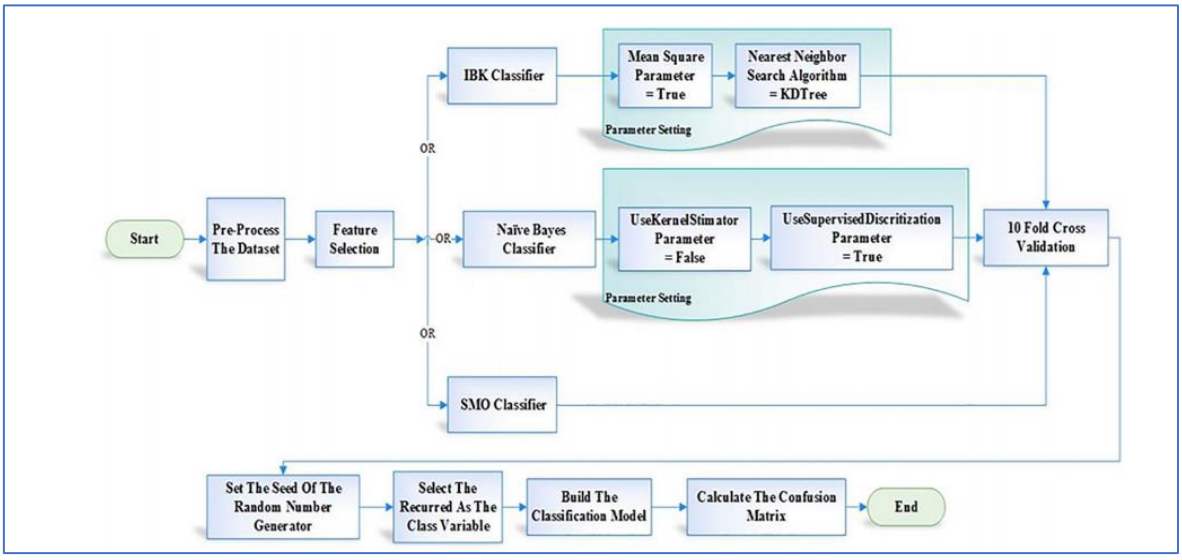

Fig. 2. The overall procedure of the decision methods.

The outcome of the review reasoned that the expectation by every one of the three calculations was worked on in the wake of preprocessing the information as far as precision and affectability. Hence, pretreatment can further develop grouping results and information quality and it ought to be noticed that the experts of RROC affirmed the legitimacy of the technique for this review led for the situation study, so this review can possibly be considered as an aide for applying the fitting pretreatment strategy to clinics all throughout the planet.

In [7] Study Entitled "Applying Data Mining Techniques to Extract Hidden Patterns about Breast Cancer Survival in an Iranian Cohort Study" (2016), this review targets making a model utilizing an information extraction calculation to extricate a significant example and data sets from a provincial dataset on bosom malignancy recuperation. The stage was the main indicator of bosom malignant growth. The information was acquired from the Research Center, an altruistic association that upholds malignancy patients in West Azerbaijan in Iran. Huge factors were removed from the paper clinical records of patients with bosom malignancy.

This review included one gathering of informational index that was dissected as the review incorporated the socioeconomics and restorative 569 patients with a normal period of 48.6 years, somewhere in the range of 2007 and 2010. The CART calculation was utilized to remove the secret example in the bosom malignancy dataset. Preceding the CART investigation, the dataset was partitioned into preparing and test information. The choice tree was worked from the preparation information, and its prescient exactness was tried by its application to foresee class naming (Survival esteems for this situation). The choice tree contains 17 hubs generally, 9 terminal hubs, and every terminal hub is related with a bunch of rules. Rules were assessed and supported by oncologists for endurance. The fundamental finding of this review was that the stage variable was the main indicator of endurance. One more aftereffect of the current concentrate additionally showed that the model made tends to foresee patients who have a higher likelihood of endurance from bosom malignant growth better than patients who have a lower probability of endurance.

With regards to [8] Study, “Comparison of Basic and Ensemble Data Mining Methods in Predicting 5-Year Survival of Colorectal Cancer Patients” (2017), it targets contrasting the proficiency of pre-judgment models dependent on numerous fundamental and expressive information in information mining techniques in foreseeing the endurance of colorectal malignant growth patients, given the significance of exact endurance expectation. CRC patients can help clinicians, scientists and medical care places to more readily anticipate patient endurance and subsequently make therapy arranging, follow-up programs and focus on medical care assets. Moreover, correlation results were acquired and can be considered by a clinical information and informatics investigation expert when giving talks on assemblages appropriate for the choice emotionally supportive network.

This study uses data from the cancer patient registry from the Digestive and Hepatology Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran. The dataset contains 1,127 records and 36 essential qualities of CRC patients who enrolled from January 2002 to 2007. The dataset is ordered through segment qualities (like age, sexual orientation, conjugal status, and societal position), indicative attributes (like starting area), and growth attributes (like histology, cancer grade), Tumor size, sickness stage (therapy and results (like a term of illness, the reason for death) utilizing meeting reports and infections put away in the malignancy scoring structure. The current review anticipated the 5-year endurance limit of CRC patients by directing a near investigation of nuts and bolts (C4.5), SVM, NB, ADTREE, RBF, REPTREE, KNN, BN, and RF) and gathering arrangement techniques. The review finishes up with an exact expectation of endurance for patients experiencing malignant growth and can uphold clinical choices and work on institutional execution in illness the board, which can be accomplished using the right information extraction calculations in settling on choice emotionally supportive networks.

With regards to [9] study entitled "Breast Cancer risk prediction using data mining classification techniques" (2015), it targets utilizing information mining procedures to group bosom disease chances as bosom malignancy represents a genuine danger to individuals' lives and is the subsequent significant reason for death for ladies and is generally normal among ladies in non-industrial nations in Nigeria where there are no administrations to help early location of bosom disease for Nigerian ladies. This review centers around utilizing two information extraction strategies to foresee bosom disease hazard in Nigerian patients utilizing choice tree and J48 calculations utilizing LASUTH patient data datasets containing hazard components and malignancy classifications (far-fetched,

plausible and harmless). The grouping of J48 choice trees for bosom disease was performed utilizing WEKA programming. The LASUTH bosom malignancy informational index was gathered from the Cancer Registry of LASUTH, Ikeja in Lagos, Nigeria. The Naive Bayes Classifier, a probabilistic model dependent on Bayes hypothesis was utilized. It is characterized as a factual classifier. It is one of the as often as possible utilized strategies for managed learning. It gives capability in the technique for managing quite a few characteristics or classes that rely upon the likelihood hypothesis. Bayesian arrangement gives calculations to functional learning and earlier information on the noticed information.

J48 choice tree is a basic choice learning calculation, the fundamental thought of ID3 is to construct a choice tree by utilizing the inquiry loop start to finish through explicit arrangements of preparing information to test every trademark without fail. It utilizes a factual property known as data gain to figure out which attribute you need to test at every hub in the tree. Data securing measures how well a given trait of a preparation test record is isolated by its order. It is reasonable for taking care of absolute just as ceaseless information. In this review, two unique arrangement strategies were utilized for information mining to foresee the danger of creating bosom disease, and malignancy hazard and execution were contrasted all together with assess the best Classifier. One of the main discoveries of this review is that the J48 choice tree is the best model for foreseeing bosom disease hazard as far as recovery exactness and precision of mistake rates recorded for the two models.

Hence, it was classified as effective in general and as an effective classifier of breast cancer risk while the quantity of qualities covered by the classifier could be expanded by expanding the example size of the preparation set and consequently fostering a more exact model.

In [10] Study entitled “Lung cancer classification using data mining and supervised learning algorithms on multi-dimensional data set” (2019), This study aims at detecting lung cancer disease through data mining. Along these lines, in this review the proposed information mining model was isolated into two unique procedures, to be executed in progression. The strategies are the arrangement technique and reasonable conglomeration displaying and accordingly, changing over malignancy information into an information base called preparing Cancer patient information. Preparing information are built with arrangement utilizing a choice tree. Order and example examination will create rehashed examples of malignant growth. A trait of the impact of the infection in a bunch called a class variable can be featured. The manner by which cellular breakdown in the lungs is analyzed is by inspecting the patient's output pictures, searching for little focuses in the lungs called knobs. Discovering a hub in itself isn't a sign of malignancy; it should contain knobs.

The issue of the review lies in the revelation of knobs in the inclusion pictures. This review exploits the Classifier organization to isolate the knob districts. These found knobs, or the components removed from them later, are utilized as contribution to the classifier to decide if a patient will foster malignant growth. The classifier likewise follows the UCI instructional exercise for knob extraction, yet one of the review downsides is that the rendition should be prepared on knobs as it was prepared on lung typical and sectioned pictures from UCI dataset and double indicators address knob areas and information on fitting lung division.

This work introduced cellular breakdown in the lungs arrangement from a UCI dataset of information extraction volumes utilizing the joined technique for unmistakable exemplary classifiers, directed learning calculations and a choice tree. The review reasoned that it is feasible to anticipate harmful growths for an extensive overview. Assuming incidentally, one hub is destructive, we presume that overall assessment is dangerous. Albeit the consequences of the knob division stage are exceptionally precise enough to anticipate the areas of nodal regions, they are now and again capricious. Knob shapes are all the more definitively characterized since some morphological cycles that produce erosive elements have been applied in the knobs. Hence, the initial phase in this methodology is to produce knob profiles to precisely ascertain the above highlights. Every one of the three calculations SVM, DT, and k-NN will in general yield knobs just as an enormous number of calculations. Along these lines, the essential advance in the grouping pipeline in this review is to diminish the bogus forecast of disease and the compelling recognition of cellular breakdown in the lungs.

Finally, in [11] Study entitled "Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study" (2019), this work aims at using data mining in breast cancer detection methods even if the data extraction classification is within the Weka Data mining program for the data set used in this study. Information mining is a factual interaction by which information is taken from an information stockroom and gathered, coordinated, and deciphered. The motivation behind information mining is to look through a lot of information to acquire significant data. Information mining characterization calculations were performed utilizing Weka. In this work, the Weka information extraction device was applied to analyze the viability of information for bosom malignancy identification. A three-dimensional bosom structure is planned comprising of the dermal layer, the lipid layer, and the sinewy glandular layer. A different model was likewise planned by including a growth the bosom structure.

Lately, options have arisen to strategies, for example, mammography x-beams, which open patients to a great deal of radiation, and bosom malignancy diagnostics have been looked for. Additionally, the point of this paper was to add to the revelation of an effortless and innocuous evaluating strategy for early determination of bosom malignancy. The high precision paces of these calculations (IBK, Random Committee, Random Forest, and Simple CART) show that bosom disease growths can really be distinguished non-intrusively for minimal price and without presenting patients to hurtful radiation, utilizing information mining grouping and hence, if the patient has malignancy bosom It is feasible to identify and analyze the cancer ahead of schedule by different strategies, for example, UWB dependent on information extraction. The arrangement strategies portrayed in this work except for customary techniques known as bosom assessments and mammograms.

3. The Research Methodology

This study used the critical evaluation method, which is based on reviewing documents and literature such as research, articles, books, etc., and examining them with criticism and evaluation to extract conclusions and indications that have relevance to answering the study questions. Therefore, this study criticized the published studies on the subject of data mining to categorize cancer diseases. In fact, that was reached by the following search tools: databases available through the Saudi Digital Library, Google Scholar, pub med search engine.

The complexities of performing an analytical task within the information analysis cycle can be described using the standard project management triangle, that is, it is necessary to complete the task and present the outcome under three main constraints: budget, timing and scope of work. In many cases, these three limitations compete with each other: in the standard information analysis task, an increase in the workload requires an increase in time and budget; A tight deadline will likely mean a budget increase and simultaneous reduction in workload, and a tight budget will likely mean less work and a shrinking project timeframe. The emergence of bottlenecks in the process of analyzing information usually leads to significant friction in carrying out the research task within the cycle of information analysis in the early stages of developing a program for this analysis. Since resources are limited, the most important bottlenecks must be addressed first. Does the information analysis team have sufficient capacity to do this? Do you need additional training? Or is it a problem that analysts lack valuable information to work with - in other words, the most important obstacle is information gathering? Or perhaps the information analysis team simply does not have the time, i.e. the group is unable to respond in time to urgent requests?

There are two ways to improve the efficiency of an analytical task within the information analysis cycle. The "productivity" of the session, that is, the accuracy with which the information analysis team can handle the analytical tasks at each stage, and the speed with which the question is answered. In fig. Figure 2 illustrates the difference between these approaches and, in general, the difference between strategic analysis tasks and queries that require a rapid response.

4. Results And Discussions

Through the analysis of previous studies, we found that the studies agree in explaining the important role of the data mining technique for classifying cancer diseases, especially after several algorithms were put under observation and tracked the results, which resulted in their overall positive role in properly supporting the health field.

Studies (6), (7), (9) and (11) agreed in searching in data mining to categorize breast cancer, but each study differs in the algorithms used (Bayes, WEKA, decision tree), but they all concluded that it is predictable with breast cancer at an early stage, at a low cost, and without exposing the patient to x-rays. While Studies (3) and (4) agreed in searching for data mining to categorize stomach cancer, but they differ in terms of the method of approaching the study. Study (3) used four algorithms (Bayes, KNN, SVM, DI) and concluded that the best algorithm can be SVM as it helps doctors to detect disease through Table (1)

Table 1. Results of Classified Patients of the Gastric Cancer based on the Four Criteria of Evaluation

	Performance Measure			
Method	***F-score***	***Precision***	***Recall***	***Accuracy***
k-NN	87.17	89.47	85	87.6
NB	87.99	84.61	89.06	87.89
DT	88.37	87.69	89.06	97.89
SVM	91.99	90.78	93.24	90.08

While study (4) used (SMOIE) technique and concluded that it helps doctors pay attention to infected patients and take the necessary action as soon as possible. Studies (5) and (10) agreed in searching for data mining to categorize lung cancer, but they differ in terms of the algorithm used, so we find that Study (5) relied on a decision tree and concluded its ability to reveal treatment costs while Study (10) used several algorithms like (SVM, KNN, DI) and it concluded that the presence of cancerous nodes can be predicted and thus be judged as a malignant node. While studies (2) and (8) were unique in the type of disease. Study (2) dealt with data mining to classify liver cancer via a decision tree, while study (8) dealt with data mining to classify colorectal cancer and used several algorithms (RE, BN, KNN, SVM) as represented by the graph. It concluded that these algorithms are able to predict patient's ability to live for 5 years while he is infected.

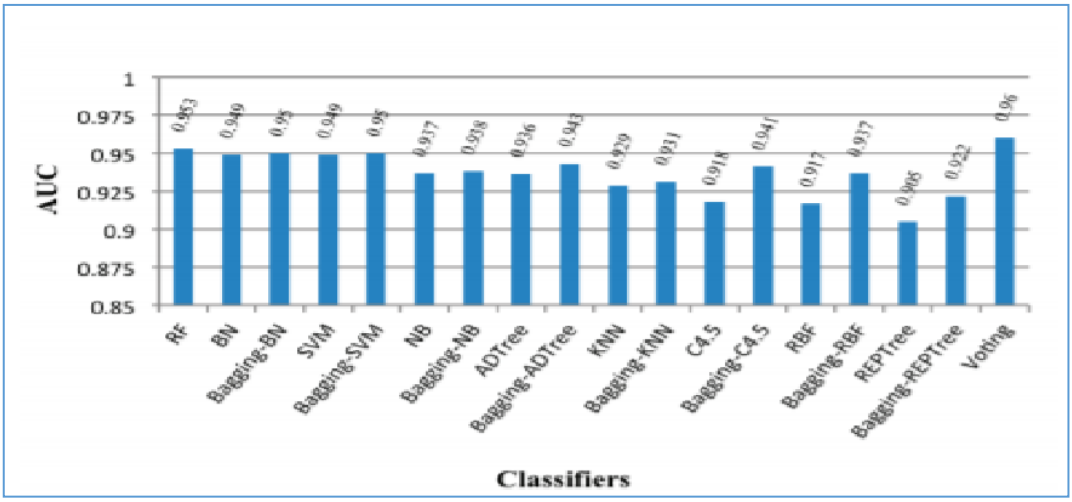

Fig. 3. Prediction performance of developed models in terms of AUC

5. Conclusion and Future Work & Directions

Data mining is a modern type of information processing technology with the huge boom in information technology, we will see a horizontal and vertical expansion in the use of data mining applications, especially in military and security applications

, Intelligence, and business, and data mining will be on mobile devices the future direction of the data. Data sets have been overwhelmed by the power of the internet Scientific, social hierarchy of data set, topology and engineering. Other features are, in particular, linking mining to graph, and analyzing social networks. Data mining faces huge databases, so it has to be algorithmic Data mining is efficient and scalable. Most current databases are Relational; Therefore, the emergence of other models of databases required the ability to process Types of data, and data mining professionals can speed up the mining process for the data, that is, an appropriate interactive interface must be provided for the beneficiaries to express Requirements and strategies; On the other hand, the interactive interface

converts Diverse outcomes for the beneficiary, i.e. a data mining system requires stronger interactivity.

The goal of this work is to clarify the technology of data mining to classify cancer diseases, and thus contribute to finding appropriate solutions by programmers and specialists, and to take advantage of the characteristics and advantages of this technology in helping doctors on the one hand and patients on the other hand through the ability of this technology to predict a state of disease severity, cost of treatment, as well as predicting the presence of disease early without surgical intervention. After review and analysis, the researcher came out with several results, the most important of which are:

- Most studies have similar results, as the algorithms of the data mining technique have proven effective in predicting the presence of disease early.
- The multiplicity of studies and foreign researches that dealt with the technology of data mining in classifying cancer diseases, but it is very few in the Arab world, and the researcher did not find Arab studies that support this type of research.
- Studies have proven the effective role of data mining techniques in reducing medical errors in terms of their ability to predict and accurately diagnose disease, but the field still needs to be further supported by applied research for new types of algorithms.
- Despite the progress made by the data mining technology, as a science, it needs more research and development in order to overcome the drawbacks resulting from its use and develop its systems to keep pace with the human mind in most of its decisions.
- By reviewing and analyzing such studies, it was found that such research is capable of developing the health system and activating the role of the patient's health file, through which the patient's health status can be predicted.

In conclusion, studies and research have proven the effective use of data mining technology into the medical side, which made robots and machine learning systems a feature of the new era as they will be relied upon and used in several medical fields. However, we find that the field needs to increase research and develop new types of algorithms and their practical application to produce the best, most accurate results that are similar to the work of doctors.

According to what we presented, the research recommended:

- Conducting intensive future studies aimed at making use of the data mining technique in predicting the severity of the disease and what its causes are.
- Hospitals adopt machine-learning systems in the medical diagnosis process.
- Conducting extensive applied research dealing with data mining
- technology in Arabic due to its absence.

In view point of forecasts for future data mining: we confirm that after the importance of the data mining process has proven in all areas of various public and private institutions, it is expected that the data mining cycle and the analytical ability to extract useful information to extract useful information is one of the most important conditions for obtaining your dream job, and you can learn more and delve into data analysis. The paid data provided by Bradford to students specializing in the fields of business administration and electronic systems, where the student or trainee is qualified to be able to master the process of data mining, which qualifies him to delve into the practical life with full force and sweep the job market successfully based on real experience and advanced study. It is expected that the data mining process will be used by all institutions in various countries of the developing world who wish to progress and achieve success, which will help raise the economic level of the country. The budget for data mining is expected to increase; Because of the correct expectations that it proved, which in turn helped many organizations to achieve success and conquer global markets, as happened with the Amazon E-Marketing Company, and the international companies and major institutions that follow. Despite the seriousness of the data mining process, which revolves around penetrating the data of users and citizens in general, with the aim of identifying the most used products and integrating them with other products to improve the sale process, cyberspace represents a real security disaster in which users are exposed to deception and theft, which makes the so-called privacy on the Internet Impossible, thus many customers and users are exposed to exploitation.

References

[1] D. L. Olson and D. Delen, *Advanced data mining techniques*. Springer Science & Business Media, 2008.

[2] J.-S. Hon, Z.-Y. Shi, C.-Y. Cheng, and Z.-Y. Li, "Applying Data Mining to Investigate Cancer Risk in Patients with Pyogenic Liver Abscess," in *Healthcare*, 2020, vol. 8, no. 2, p. 141.

[3] A. Mortezagholi, O. Khosravizadeh, M. B. Menhaj, Y. Shafigh, and R. Kalhor, "Make intelligent of gastric cancer diagnosis error in Qazvin's medical centers: Using data mining method," *asian pacific J. cancer Prev. apjcp*, vol. 20, no. 9, p. 2607, 2019.

[4] M.-M. Liu, L. Wen, Y.-J. Liu, Q. Cai, L.-T. Li, and Y.-M. Cai, "Application of data mining methods to improve screening for the risk of early gastric cancer," *BMC Med. Inform. Decis. Mak.*, vol. 18, no. 5, pp. 23–32, 2018.

[5] T. Yu, Z. He, Q. Zhou, J. Ma, and L. Wei, "Analysis of the factors influencing lung cancer hospitalization expenses using data mining," *Thorac. Cancer*, vol. 6, no. 3, pp. 338–345, 2015.

[6] Z. Sajjadnia, R. Khayami, and M. R. Moosavi, "Preprocessing breast cancer data to improve the data quality, diagnosis procedure, and medical care services," *Cancer Inform.*, vol. 19, p. 1176935120917955, 2020.

[7] H. R. Khalkhali, H. L. Afshar, O. Esnaashar, and N. Jabbari, "Applying data mining techniques to extract hidden patterns about breast cancer survival in an Iranian cohort study," *J. Res. Health Sci.*, vol. 16, no. 1, p. 31, 2016.

[8] M. A. Pourhoseingholi, S. Kheirian, and M. R. Zali, "Comparison of basic and ensemble data mining methods in predicting 5-year survival of colorectal cancer patients," *Acta Inform. Medica*, vol. 25, no. 4, p. 254, 2017.

[9] K. Arutchelvan and R. Periyasamy, "Cancer prediction system using datamining techniques," *Int. Res. J. Eng. Technol.*, vol. 2, no. 08, 2015.

[10] S. R. A. Ahmed, I. Al Barazanchi, A. Mhana, and H. R. Abdulshaheed, "Lung cancer classification using data mining and supervised learning algorithms on multi-dimensional data set," *Period. Eng. Nat. Sci.*, vol. 7, no. 2, pp. 438–447, 2019.

[11] M. K. Keleş, "Breast cancer prediction and detection using data mining classification algorithms: a comparative study," *Teh. Vjesn.*, vol. 26, no. 1, pp. 149–155, 2019.