8-10-2020

Data Mining Algorithms Predicting Different Types of Cancer: Integrative Literature Review

Ahmad Al-Aiad
Jordan University of Science and Technology, aiaiad@just.edu.jo

Salsabil Abualrub
Jordan University of Science and Technology, salsabil.abualrub@just.edu.jo

Yazan Alnsour
University of Northern Colorado, yazan.alnsour@unco.edu

Mohammad Alsharo
Al Albayt University, mohammad.alsharo@aabu.edu.jo

Data Mining Algorithms Predicting Different Types of Cancer: Integrative Literature Review

Abstract

Based on the World Health Organization, cancer is the second leading cause of death globally and is responsible for an estimated 9.6 million deaths in 2018. Globally, about 1 in 6 deaths is due to cancer, and approximately 70% of deaths from cancer occur in low and middle-income countries. with accelerating developments in technologies and the digitization of healthcare, a lot of cancer's data have been collected, and multiple cancer repositories have been created as a result. cancer has become a data-intensive area of research over the last decade. A large number of researchers have used data mining algorithms in predicting different types of cancer to reduce the cost of tests used to predict different types of cancer, especially in low and middle-income countries. This paper reports on a systematic examination of the literature on data mining algorithms predicting different types of cancer through which we provide a thorough review, analysis, and synthesis of research published in the past 10 years. We follow the systematic literature review methodology to examine theories, problems, methodologies, and major findings of related studies on data mining algorithms predicting cancer that were published between 2009 and 2019. Using thematic analysis, we develop a research taxonomy that summarizes the main algorithms used in the existing research in the field, and we identify the most used data mining algorithms in predicting different types of cancer. In addition, to data mining algorithms used in predicting each type of cancer, as mentioned in the reviewed studies. We also identify the most popular types of cancer that researchers tackled using predictive analytics.

Key Words: Data Mining Algorithms, Predicting Cancer, Cancer Types, Integrative Literature Review

Introduction

As increase of availability of quality data on different fields, the integration of data repositories into warehouses and the increase in data processing and storage capabilities in parallel with decrease cost, the concept of data mining emerged. Data mining is the process that uses statistical, mathematical and artificial intelligence techniques to extract and identify useful information and knowledge (patterns) from large sets of data involving methods at the intersection of machine learning, statistics, and database systems. Once the information and patterns are found it can be used to make decisions for developing the business [1]-[5]. Data mining tools can give answers to your various questions related to your business which was too difficult to resolve, other names associated with data mining: knowledge extraction, pattern analysis and knowledge discovery. Based on the way in which the patterns are extracted from the historical data, the learning algorithm can be classified as supervised or unsupervised. With supervised learning algorithms, the training data includes both the descriptive attributes and the class attributes. with unsupervised learning the training data includes only the descriptive attributes. The data mining consists of various Techniques some of these techniques are tracking patterns, outlier detection (Identifying Anomalies) and association [8][9].

Data mining used in healthcare; Data mining holds great potential to improve health systems. It uses data and analytics to identify best practices that improve care and reduce costs. Mining can be used to predict the volume of patients in every category. Generally, it discovers the relationships between diseases and the effectiveness of treatments. That is to identify new drugs or to ensure that patients receive appropriate, timely care [13]. data mining can also help healthcare insurers to detect fraud and abuse. Data mining helps identify the patterns of successful medical therapies for different illness. Classification is a data mining technique that assigns categories to a collection of data in order to aid in more accurate predictions and analysis, classification is the most frequently used data mining technique for real world problems (e.g. healthcare problems) it's part of the machine-learning family, classification employ supervised learning for example we use classification to predict the type of disease given symptoms [18] .The most field that use classification technique is healthcare, data mining classification techniques applied for breast cancer diagnosis and prognosis.

In the last few years, a significant amount of research has been conducted on predicting cancer using data mining algorithms. the main purpose of this research is to systematically review, analyze, and synthesize existing research on predicting different types of cancer using data mining algorithms, it provides a research taxonomy and framework which offer important insights for researchers working on data mining algorithms to predict the presence of different types of cancer. We aim to answer the following research questions.

RQ1.What is the mostly used data mining algorithms in predicting different types of cancer?

RQ2.What is the data mining algorithms used in predicting each type of cancer?

RQ3.What is the most popular type of cancer that researchers have made studies on it and used data mining algorithms in predicting it?

To answer these research questions, we conducted a systematic literature review to identify high-quality research papers that used data mining algorithms in predicting cancer. To this end, we followed the systematic literature review methodology proposed by IEEE transactions on professional communication journal. In their methodology, the systematic review process consists of three phases: planning the review, conducting the review, and data extraction and synthesis.

The rest of this paper is organized as follows. First, the systematic literature review process employed in this research is discussed. Second, an analysis and discussion of the results is provided. Third, conclusions.[22]

RESEARCH METHODOLOGY: INTEGRATIVE LITERATURE REVIEW PROCESS

the systematic review process consists of three phases: planning the review, conducting the review, and data extraction and synthesis. Fig. 1 shows the three phases of the methodology and their steps.[22]

In planning the review, the aim is to specify a topic to be covered and the research questions to be addressed. In this study, the researchers decided to review published literature on predicting cancer using data mining algorithms to identify the mostly used algorithms in researches on the topic.

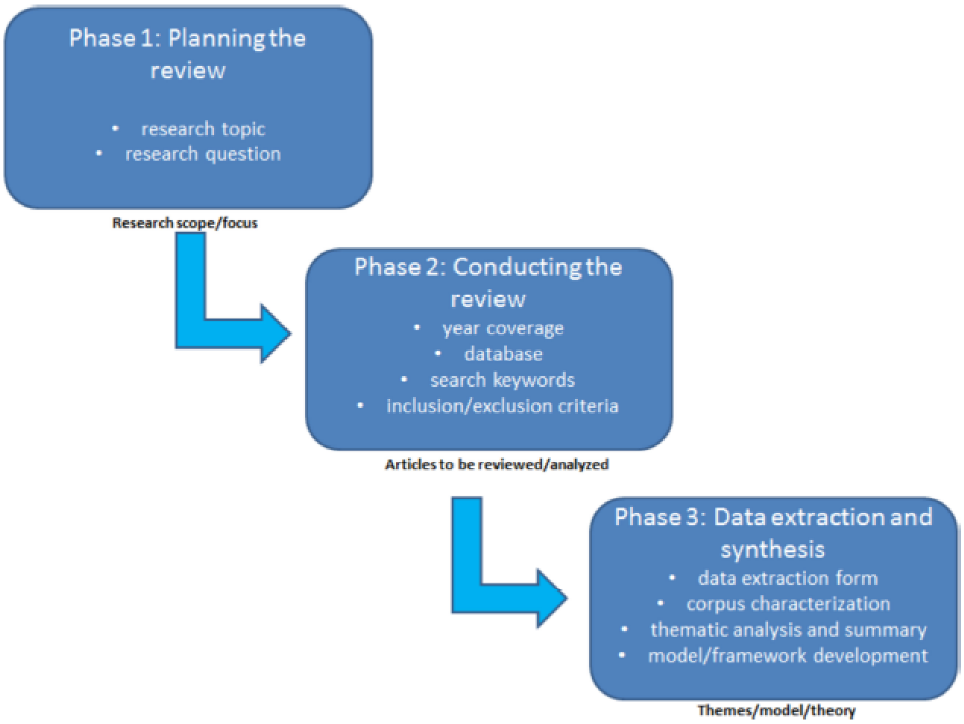
Phase 1: Planning the review
• research topic
• research question
Research scope/focus
Phase 2: Conducting the review
• year coverage
• database
• search keywords
• inclusion/exclusion criteria
Articles to be reviewed/analyzed
Phase 3: Data extraction and synthesis
• data extraction form
• corpus characterization
• thematic analysis and summary
• model/framework development
Themes/model/theory

Fig. 1. Research methodology: The systematic review process.

The goal of the second phase, conducting the review, is to conduct a comprehensive, unbiased search on the literature, based on numerous search rules and parameters, to identify a set of articles to be reviewed. These rules and parameters include year coverage, database to be searched, keywords and search terms, and inclusion and exclusion criteria. The process begins by identifying the span of research publication years to be covered. To reflect the current state of the research, we include research papers published between 2009 and 2019.

As a search strategy, to acquire a handful of quality sources for the review, the researchers decided to conduct the search in three of the most recognized databases: Medline, Pubmed, IEEE Xplore. The next step was to define rules to govern the literature search, identifying specific, relevant search terms and keywords to be used in the research retrieval process. The search terms and keywords that we used during initial search are data mining predict cancer, cancer detection and data mining and data mining detect cancer.

We applied the advanced search features in ways that allowed us to maintain consistency across all databases. For our initial review, we searched in the abstract, keyword, and title. We excluded all duplicates, based on the title and abstract review, to yield an initial corpus of 495 published works used for further analysis.

The next step was to formulate inclusion and exclusion criteria. Through an iterative process, the researchers came up with the following criteria.

1. The corpus should include only studies published in English Studies published in languages other than English were excluded (64 studies were excluded).

2. The corpus shouldn't include articles that did not undergo peer review process according to journal policy (145 studies were excluded).

3. The corpus should also include only journal papers on conference. Entries in the initial corpus, including books, book chapters, notes, and technical reports, were excluded. (5 studies were excluded).

4. Of the remaining articles and proceedings papers, we reviewed the titles and abstracts using our selection rules related to the content, described below. This left us with a set of 83 articles and proceedings (198 studies were excluded).

5. Of the remaining set of articles and proceedings, we reviewed the full text using our selection rules related to the content: 63 were excluded, and 20 articles and proceedings papers were used for further analysis through data extraction and synthesis.

For our selection rules related to the content, we used three rules for inclusion.

1. We included articles and proceedings papers that build data mining algorithms Used real and reliable data set.

2. We included articles and proceedings papers that used Data mining tools and algorithms to present the main role in predict the presence of different types of cancer.

3. We included articles and proceedings papers with high accuracy and efficiency of used algorithms.

Fig. 2 shows the size of the corpus and the effects of applying the inclusion and exclusion criteria at each stage.

We conducted the research selection process in several rounds, ultimately uncovering a total of 20 research articles and conference proceedings papers that matched our research criteria. Then, we proceeded to the next step of the systematic review process: data extraction and synthesis. The main goal of data extraction is to examine the items in the final corpus and record their features of interest as related to the original research topic and questions. The main goal of data synthesis is to summarize and integrate the findings of the final corpus of articles reviewed. The next section will describe the data extraction and synthesis stage in detail.

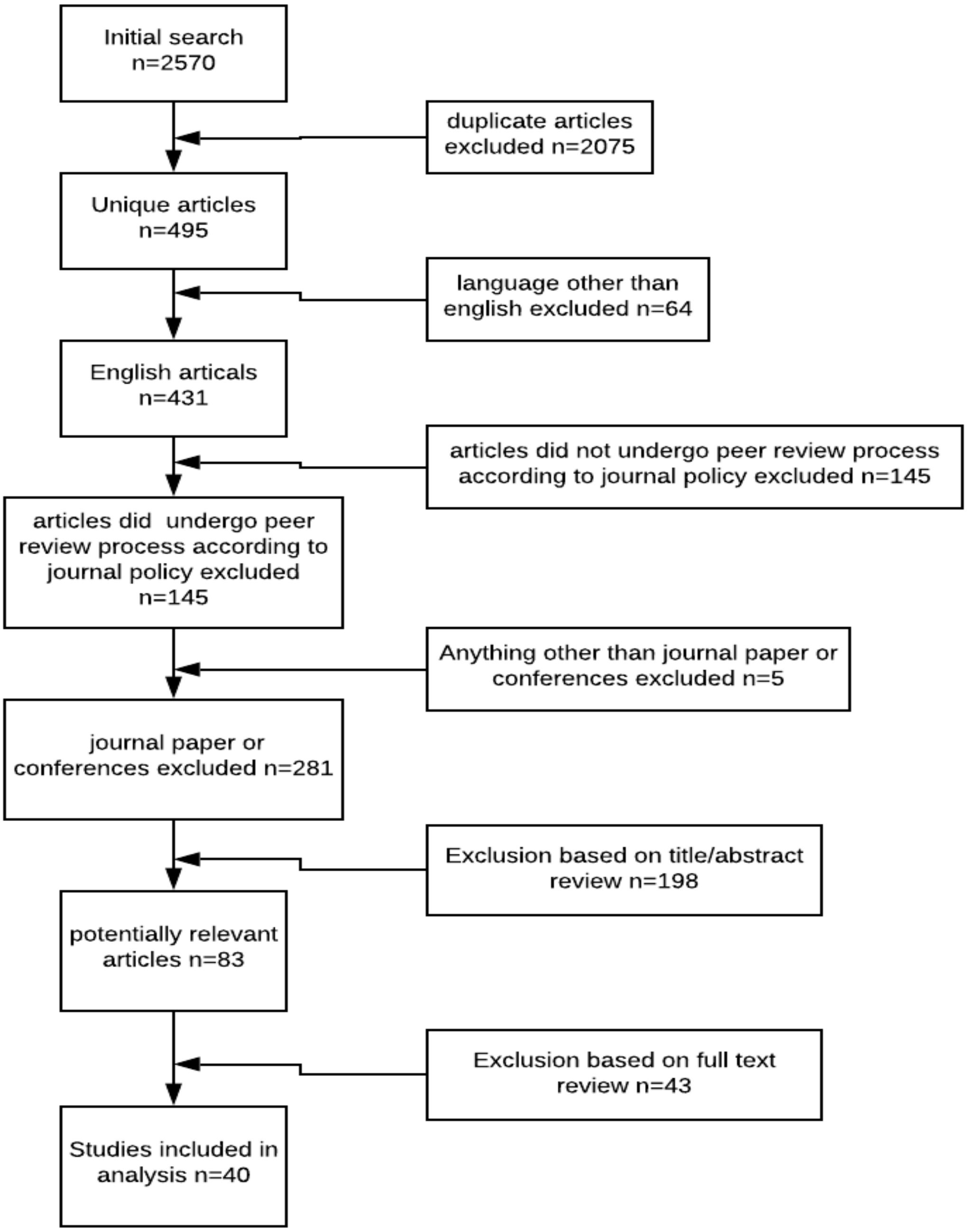
Initial search n=2570
duplicate articles excluded n=2075
Unique articles n=495
language other than english excluded n=64
English articals n=431
articles did not undergo peer review process according to journal policy excluded n=145
articles did undergo peer review process according to journal policy excluded n=145
Anything other than journal paper or conferences excluded n=5
journal paper or conferences excluded n=281
Exclusion based on title/abstract review n=198
potentially relevant articles n=83
Exclusion based on full text review n=43
Studies included in analysis n=40

Fig.2. Number of articles in each stage based on inclusion and exclusion criteria.

ANALYSIS AND RESULTS: DATA EXTRACTION AND SYNTHESIS

After selecting a corpus of 20 articles and conference proceedings based on the inclusion and exclusion criteria described above, we proceeded to the last phase of the systematic review process: data extraction and synthesis. To achieve the research goal and provide answers to the research questions mentioned earlier (i.e., RQ1, RQ2, and RQ3). Through an iterative process, we identified the following features of the data extraction form: Data set used, the role of the algorithm, algorithms accuracy, used algorithms, major findings, method, objective, solved problem.

Taxonomy Development

We developed a thematic taxonomy of research (see Fig. 3). The research focused on used data mining algorithms in predicting different types of cancer. The taxonomy's main purpose is to provide a comprehensive reference model and road map that helps researchers understand the algorithms used in predicting each type of cancer.

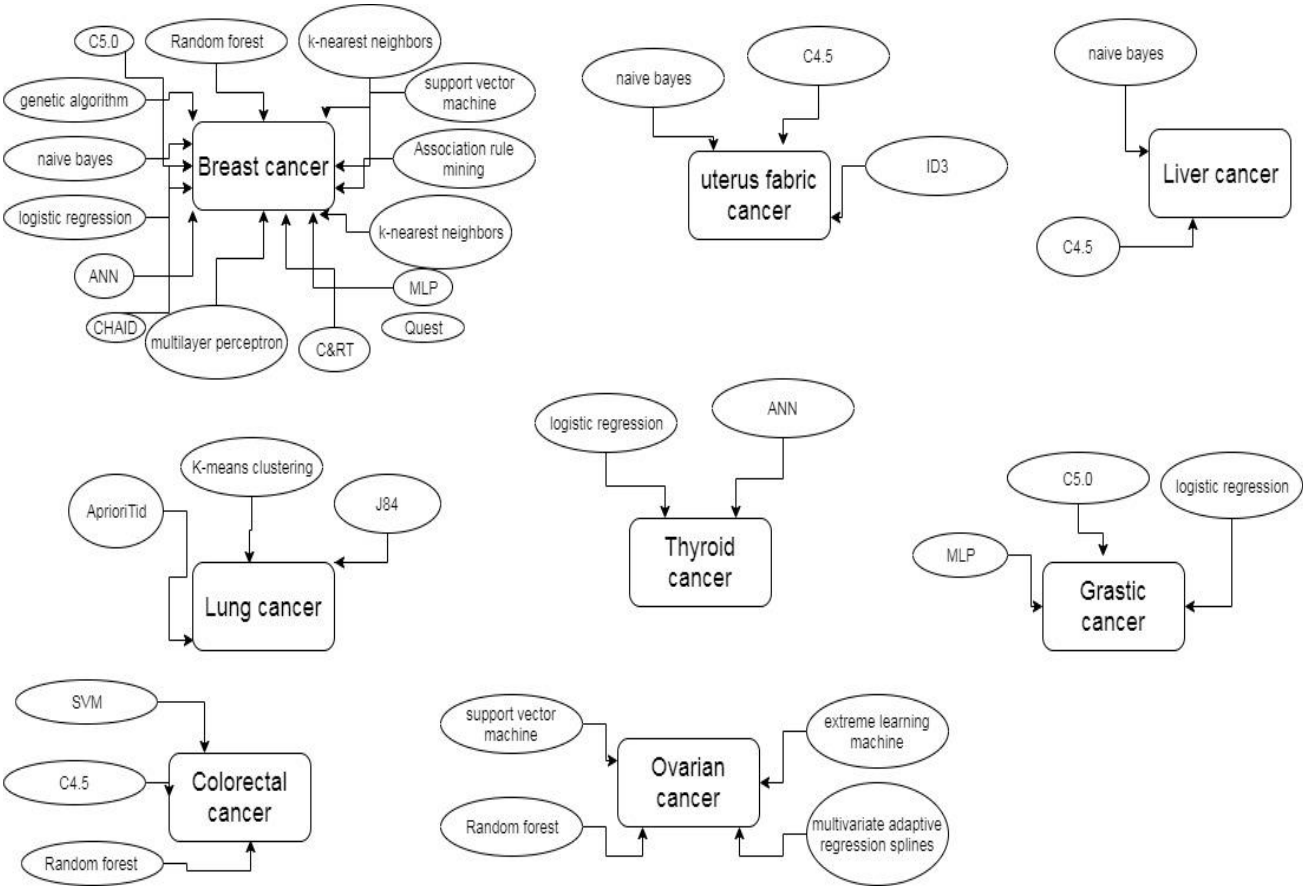

Fig. 3. Taxonomy of research

The taxonomy represents 8 types of cancer which is breast cancer, lung cancer, liver cancer, gastric cancer, colorectal cancer, thyroid cancer, ovarian cancer and uterus cancer, and the data mining algorithms used in predict each type of cancer.

Results/Discussion

The researchers made studies on 8 types of cancer which is breast cancer, lung cancer, liver cancer, gastric cancer, colorectal cancer, thyroid cancer, ovarian cancer and uterus cancer. most studies are on breast cancer; this is because breast cancer is a major cause of concern in the USA today. For instance, it affects one in every seven women in the USA; that's why most researchers focus on using data mining algorithms to predict breast cancer. different algorithms are used to predict breast cancer such as Random forest ,k-nearest neighbors, support vector machine, Association rule mining, k-nearest neighbors, MLP, Quest, C&RT, multilayer perceptron, CHAID, ANN, logistic regression, naive bayes, genetic algorithm and C5.0, The mostly used algorithm in predicting breast cancer is decision tree.(J48, C4.5, C5.0, ID3 and Random forest all are types of decision tree). Then liver cancer and lung cancer are the second mostly types of cancer that researchers make studies on, for liver cancer they have used naive bayes and C4.5 algorithms. For lung cancer they have used different types of decision tree, K-means clustering and AprioriTid. [1]-[20]

Decision tree algorithm is the mostly used algorithm in predicting different types of cancer, literally It has been used to predict 7 types of cancer which is breast cancer, lung cancer, liver cancer, gastric cancer, colorectal cancer, thyroid cancer, ovarian cancer and uterus cancer; that's because It is easy to understand and it is one of the most popular data mining classification algorithms also there is multiple decision tree types (such as ID3, C4.5, CART, CHAID, MARS, and Conditional Inference Trees) used in different data mining tools which make it flexible and easy to find in all data mining tools such as Weka, Tanagra, Rapidminer and others. Moreover, this algorithm attained high level of accuracy in all studies that used it, this result in encourage researchers to use decision tree algorithm on their studies.[1]-[20]

Figure 4 represents the Decision tree algorithm accuracy rate in predicting seven types of cancer:

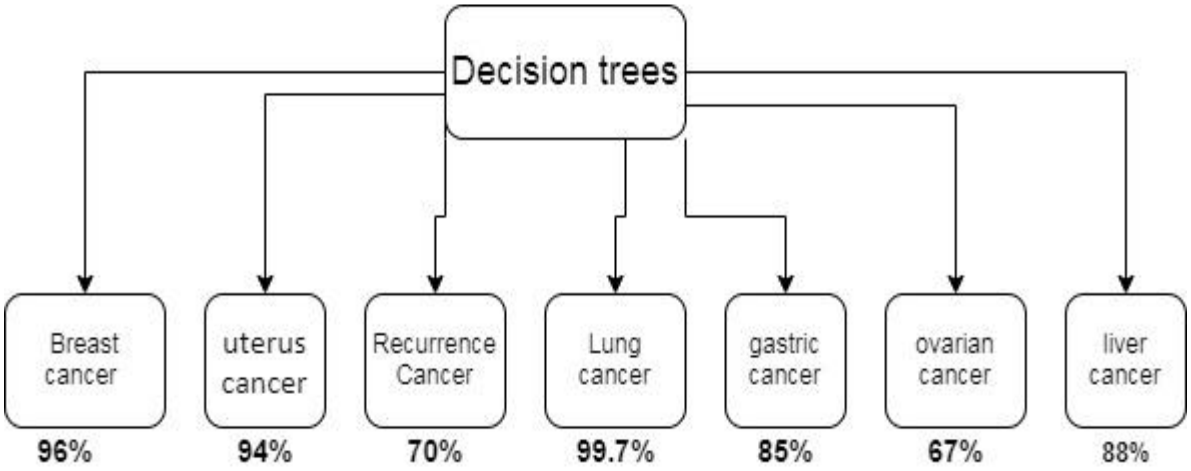

Fig.4. Decision tree accuracy

Conclusions

The primary aim of our study has been to conduct a systematic literature review on predicting different types of cancer using data mining algorithms with the intention of answering the three research questions RQ1, RQ2, and RQ3. We have developed a taxonomy of 7 major cancer types and the algorithms used to predict each type. This study represent a guideline for future research ,it help researchers know which is the best algorithms to use in their study based on the type of cancer they will make the study on and this can reduce time and effort consumed on try huge number of algorithms to find the most suitable algorithms to use in their studies.

References

(1) S. Bharati, M. A. Rahman, and P. Podder, “Breast Cancer Prediction Applying Different Classification Algorithm with Comparative Analysis using WEKA,” 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT), 2018.

(2) N. Ramkumar, S. Prakash, S. A. Kumar, and K. Sangeetha, “Prediction of liver cancer using Conditional probability Bayes theorem,” 2017 International Conference on Computer Communication and Informatics (ICCCI), 2017.

(3) D. K. S. Girija and M. S. Shashidhara, “Data mining techniques used for uterus fibroid diagnosis and prognosis,” 2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 2013.

(4) N. Bhargava, S. Sharma, R. Purohit, and P. S. Rathore, “Prediction of recurrence cancer using J48 algorithm,” 2017 2nd International Conference on Communication and Electronics Systems (ICCES), 2017.

(5) M. V. Dass, M. A. Rasheed, and M. M. Ali, “Classification of lung cancer subtypes by data mining technique,” Proceedings of The 2014 International Conference on Control, Instrumentation, Energy and Communication (CIEC), 2014.

(6) U. D. R and B. Ramachandra, “Association rule mining based predicting breast cancer recurrence on SEER breast cancer data,” 2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), 2015.

(7) Q. Fan, C.-J. Zhu, and L. Yin, “Predicting breast cancer recurrence using data mining techniques,” 2010 International Conference on Bioinformatics and Biomedical Technology, 2010.

(8) Q. Fan, C.-J. Zhu, and L. Yin, “Predicting breast cancer recurrence using data mining techniques,” 2010 International Conference on Bioinformatics and Biomedical Technology, 2010.

(9) M. Jajroudi, T. Baniasadi, L. Kamkar, F. Arbabi, M. Sanei, and M. Ahmadzade, “Prediction of Survival in Thyroid Cancer Using Data Mining Technique,” Technology in Cancer Research & Treatment, vol. 13, no. 4, pp. 353–359, 2014.

(10) 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 Medical Informatics and Decision Making, vol. 18, no. S5, 2018.

(11) J. Diz, G. Marreiros, and A. Freitas, “Applying Data Mining Techniques to Improve Breast Cancer Diagnosis,” Journal of Medical Systems, vol. 40, no. 9, Jun. 2016

(12) C.-J. Tseng, C.-J. Lu, C.-C. Chang, G.-D. Chen, and C. Cheewakriangkrai, “Integration of data mining classification techniques and ensemble learning to identify risk factors and diagnose ovarian cancer recurrence,” Artificial Intelligence in Medicine, vol. 78, pp. 47–54, 2017

(13) M. Pourhoseingholi, S. Kheirian, and M. Zali, “Comparison of Basic and Ensemble Data Mining Methods in Predicting 5-Year Survival of Colorectal Cancer Patients,” Acta Informatica Medica, vol. 25, no. 4, p. 254, 2017

(14) K. Ahmed, A.-A.-E. Abdullah-Al-Emran, T. Jesmin, R. F. Mukti, M. Z. Rahman, and F. Ahmed, “Early Detection of Lung Cancer Risk Using Data Mining,” Asian Pacific Journal of Cancer Prevention, vol. 14, no. 1, pp. 595–598, 2013

(15) D. Delen, G. Walker, and A. Kadam, “Predicting breast cancer survivability: a comparison of three data mining methods,” Artificial Intelligence in Medicine, vol. 34, no. 2, pp. 113–127, 2005

(16) A. Çakır and B. Demirel, “A Software Tool for Determination of Breast Cancer Treatment Methods Using Data Mining Approach,” Journal of Medical Systems, vol. 35, no. 6, pp. 1503–1511, Feb. 2010

(17) S. Shah and A. Kusiak, “Cancer gene search with data-mining and genetic algorithms,” Computers in Biology and Medicine, vol. 37, no. 2, pp. 251–261, 2007

(18) P. J. García-Laencina, P. H. Abreu, M. H. Abreu, and N. Afonoso, “Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values,” Computers in Biology and Medicine, vol. 59, pp. 125–133, 2015

(19) M. Ring and B. M. Eskofier, “Data Mining in the U.S. National Toxicology Program (NTP) Database Reveals a Potential Bias Regarding Liver Tumors in Rodents Irrespective of the Test Agent,” Plos One, vol. 10, no. 2, Jun. 2015

(20) G. Ge and G. W. Wong, “Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles,” BMC Bioinformatics, vol. 9, no. 1, Nov. 2008

(21) H. L. Afshar, M. Ahmadi, M. Roudbari, and F. Sadoughi, “Prediction of Breast Cancer Survival Through Knowledge Discovery in Databases,” Global Journal of Health Science, vol. 7, no. 4, 2015

(22) A. Alaiad, Y. Alnsour and M. Alsharo, "Virtual Teams: Thematic Taxonomy, Constructs Model, and Future Research Directions", *IEEE Transactions on Professional Communication*, vol. 62, no. 3, pp. 211-238, 2019. Available: 10.1109/tpc.2019.2929370.

(23) "Home", *Who.int*, 2020. [Online]. Available: https://www.who.int/. [Accessed: 10- Jan- 2020].