ISSN: 2582-8118 ***Volume 2, Issue 1; Jan 2022***

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYSIS IN SCIENCE AND ENGINEERING

Web: https://www.iarj.in/index.php/ijrase/index

3. A Systematic Review of Data Mining in Health Care: A Case of Breast Cancer

Ibrahim Hassan
Department of Computer Science,
Gombe State Polytechnic, Bajoga.

E. J. Garba, A. S. Ahmadu
Department of Computer Science,
Modibbo Adama University, Yola.

ABSTRACT

Breast cancer is a major health issue that affects women all over the world recording about 23% of all malignancies. The disease was the leading cause of female mortality in developed countries, second in the world, and third in developing countries, amounting to 14% of cancer fatality. Data mining is an excellent tool in data science that allows practical scenarios to be analysed using various techniques. In the field of medicine, analysing records with large data become a herculean task and time pressing issue. Thus, the most challenging aspect in this area is not just to identify that a breast tumor is benign or malignant but also knowing the stage of the disease as early as possible, this will help in treatment and prognosis of the disease. In this field, several research studies are now underway. As a result, a careful examination of completed and unfinished research is required in order to open up a new study horizon. Hence, the goal of the study is to provide an in-depth review of data mining application in health care, with focus on breast cancer.

KEYWORDS

Data mining, benign, malignant, prognosis, breast cancer stage.

Introduction:

Data mining is a new technology and has successfully applied in many fields (Nalii & Meera, 2018). The overall goal of data mining process is to extract information from a data set and transform it into an understandable structure for further use and is mainly used for classification and prediction (Agrawal & Gupta, 2013). Data mining, is a pattern discovery and extraction technique that involves a large amount of data. For clinical and prognosis purposes, data mining and health care area have implemented some detection systems as well as other health care related technologies (Ahmed, Laila, Sherif, & Ayman, 2018).

To find the useful and hidden knowledge from the database is the purpose behind the application of data mining (Iffat, Ankita, Nandan, & Bindu, 2020). Popularly data mining called knowledge discovery from the data. Data Mining has been used in a variety of applications such as Marketing, Customer Relationship Management, Engineering, and Medicine analysis, Expert Prediction, Web mining, and Mobile computing (Haifeng & Sang Won, 2015).

Data mining algorithms applied in healthcare industry play a significant role in prediction and diagnosis of diseases (Preeta & Vinila, 2019). There is a large number of data mining applications found in the medical related areas such as Medical device industry, Pharmaceutical Industry and Hospital Management (Durairaj & Ranjani, 2013).

Today's health industry creates a vast amount of complicated data regarding patients, hospital resources, diagnosis of diseases, electronic database of patients, and various medical gadgets, among other things (Prabakaran & Jagadeesh, 2016). Data mining applications are being developed to help with health care management by evaluating the success of medical treatments, identifying and tracking chronic illness states, designing suitable interventions, and reducing the hospital admissions and claim (Rozita, Nasroalla, & Saeid, 2017).

The beauty and success of any research is to obtain a reliable data. The data bank contains vital hidden knowledge which could be generated through the use of data mining technique (Padapriya & T. Velmarugan, 2014) However, series of researches have and are still taken place to overcome the challenges of delivering high quality in the health sector, particularly in breast cancer diseases.

Breast cancer is a major health issue that affects women all over the world, recording 23% of all malignancies, it equally resulted to about 14% of cancer fatalities (Hamza & Nagendra, 2020).

Breast cancer is a disease that results from the uncontrolled growth of cells in the breast. The condition primarily affects women, though it can sometimes affect men (Prabakaran & Jagadeesh, 2016).

2.0 Data Mining Tools Used in Healthcare Databases:

Cela & Frasheri (2013) analyses different data mining techniques and tools used in health care databases with the aim to identify the most important application fields and trends of research in health care domain. Method adopted in the research include: review of journals of Computer science, Engineering, and Health care. They identified that health care databases have huge amount of data but lack of effective analysis tools to discover the hidden knowledge with appropriate computer based information and decision support system that will help physicians and other health workers suggest less expensive therapeutically equivalent alternatives. Application of data mining technique is not restricted to diagnosis, prognosis, and treatment but also to predict result of surgeries. The research categorized the data mining techniques into: classical and new algorithms. The new algorithm is further sub divided into: novel and hybrid. An algorithm from each of these sub division was implemented and tested on cardiovascular diseases and prostate

cancer. Results obtained show that MLPNN algorithm recorded 100% accuracy when subjected to 15 attributes with 573 data instances in diagnosing and prediction of cardiovascular diseases. Next is decision tree with 99.62% accuracy when 15 attributes were subjected on 573 data instances. The least, been KNN with 45.67% o on 14 attributes with 3,000 instances of data. While in cancer disease, decision tree (C4.5) algorithm recorded 100% accuracy on breast cancer and 90% on prostate cancer. The least, been logistic regression with 89.61% on prostate cancer. Their research, analyses that most of the researches done, were mainly focused in studying classic data mining algorithms in which decision tree is one of them and thus, showing acceptable level of accuracy. Future research, will aim at dealing with data mining technique that have wider spectra of application for group of diseases with more attention on treatment options for patients.

However, Sowmiya, Gopi, Begin, & Robinson (2014) attempted to optimize lung cancer using modern data mining technique by suggesting an improved version of Intensity Modulated Radio Therapy (IMRT) called ROCO (Reduced Order Constrained Optimization) to be applied to difficult behaviour sites (like the lung). ROCO was previously applied to prostate cancer but their work of 2014 focuses on treatment for in operable, non-small cell lung cancer (NSCLC) with the aim of providing a better system that will reduce the amount of time required to obtain a clinically main acceptable IMRT plans as well as to enable medical treatment planners to focus on critical trade-offs between tumor coverage and normal body organs. Above all, is to mine lung cancer data to discover knowledge that is never to be only correct but also comprehensible for lung cancer detections. Some of the problems attributed to IMRT that pave a way for the design of ROCO has to do with its long planning times that place a severe stress on available resources in a busy medical clinic (which can result to a treatment delays, acceptance of a neighbour sub optimal plans or in the worst case, errors due to the time pressure). The design of ROCO, centrally rely on Ant Colony Optimization (ACO) algorithm. Though solutions offered by this very algorithm may be far from optimal and choices made at early stages reduce a set of possible steps at latter stages, but ACO was proposed in order to model the problem at hand to mimic the ant system. Most significantly, ROCO uses the clinical full dose distributions corresponding to each PCA method, using an unestimated result in an inaccurate dosage scheming. Implementation is done with an integral part of the medical clinical MSKCC (scheduling system) and treatment planning system in order to make it feasible and stretchable enough to deal with the different type of treatment sites beside the prostate gland. In addition, the IMRT is improved given rise to ROCO in such a way that, when beam is projected to the image slice it separate component into different colours. Results obtained from their work show that, acceptable plans could be obtained in approximately 30 minutes. ROCO strategies satisfy all of the clinical restrictions imposed by the medical plans to an extent that there were no significant differences between the OAR sparing achieved by ROCO and the organ sparing scheduling achieved by medical plans. Future focus of the research, will apply searching algorithm different from ACO to mine cancer data and fid better grouping and correctness and determine performance analysis for such algorithms. Also, future work will apply ROCO to skull and neckline cancer which remains a puzzling site for current IMRT planning technique.

The work of Luo, Houcberg, & Uzoner (2015) formulated a system known as Sub graph Augmented Non-negative Tensor Factorization (SANTF) after observing that transformation from clinical cases and experiences to knowledge is largely an expert task,

periodic labour intensive and reports are however attributed to errors. The goal of the research includes design of an automated approach that will keep track of records and give a clear interpretation of (verbal) reports with zero or minimal bias. Also assist experts to make review and cover a large patient population. The system outlined a workflow of a narrative text sentences converted to graph derived from natural language processing for pathology reports. Frequent sub graph mining tools are used to collect sub graphs. Finally, sentences are grouped to identify association between the captured sentences. Results demonstrated over 10% improvement in averaged f-measure on patient clustering compared to the widely used non-negative matrix factorization. Main problem of the system is lack of a dedicated case study, as health institutions have different pattern and style of documentation.

Vijaya, Suhasini, & Priya (2014) made a study and used different classification algorithms (Naive Bayes, Support vector machine, k-Nearest Neighbour, and J48 to analyse lung cancer disease prediction. Analysis of these classification algorithm was done using WEKA tool. Classification accuracy is generally calculated by the percentage of instances it has correctly classified. The output includes the result of the data values given in the data set. The results obtained showed that all the algorithms performed significantly better performance. Their study does not take care of pinpointing a single classifier to be better than the other. Though there are chances to obtain results that are of varying performance even in a similar dataset. However, the study can be extended by applying additional range of classification algorithms or use a proposed algorithm on additional datasets from huge medical databases from various domains.

3.0 Applications of Data Mining Techniques in Analysis of Breast Cancer:

Gupta, Kumar, & Sharma (2011), conducted a research by making a review on various technical articles of breast cancer diagnosis and prognosis. Six (6) different data mining classification methods: decision tree, support vector machine, genetic algorithm, fuzzy sets, neural networks and rough sets were analysed under the review for the diagnosing while for the prognosis, the American Joint Commission on cancer (AJCC) staging system based on TNM (Tumour, Node, Metastasis) staging is used with Artificial Neural Network (ANN) algorithm. Both experiments were carried out using WEKA and SEER for the analysis.

Their review offers great promise to uncover patterns hidden in the data that can help in decision making, so also, the prognostic problem is mainly analysed under ANN and its accuracy came higher in comparison to other classification techniques applied for same. Thus, their study recommends more efficient models to be provided using different technologies and algorithms.

Arafat, Barakat, & Goweda (2012), also presented a research with the aim of using intelligent techniques for breast cancer classification. The research investigated a strategy based on Rough set theory with particle swarm optimization to achieve intelligent solutions for complex problems by proposing a hybrid approach that can help in improving classification accuracy and also in finding more robust features to improve classifier performance. Their research recommends the blending process to combine the advantages of particle swarm optimization with the advantages of other intelligent optimization algorithms to create a compound algorithm that has practical value.

In addition to this, is the research of Kharya (2012), where various data mining approaches that have been utilized for breast cancer diagnosis and prognosis were discussed. The research, analyses two (2) classification techniques using association rule mining and ANN to detect and classify breast cancer in digital mammography. The result shows that the two approach performed well, recording classification accuracy of over 70% for both techniques. The research constructed a decision tree model with a small sample size, instead a large data set should be used to build high degree of statistical confidence. He pointed out the research of Bllachia and Erhan which analyses three (3) different data mining technique:

Naive Bayes, Artificial Neural Network (ANN) and C4.5 decision tree algorithm to predict the survivability rate of breast cancer. The analysis was carried out using WEKA toolkit and SEER with 485 records and finds out that all the techniques have accuracy of more than 80% with decision tree having the highest (86.7%). The research recommends future study to look into the development of computer aided diagnostic tools. Among the techniques pointed out used in breast cancer diagnosis, decision tree is found to be the best predictor with 93.62% accuracy on benchmark data set and SEER. He further recommends future work to focus on using the best predictor to design a web based application, as such can be implemented in remote areas to imitate human like diagnostic expertise for prediction.

Mumine (2019), presented a study whose purpose was to predict and detect breast cancer in its ealier stage even if the tumor size is smaller. As such, comparative analysis of data mining classification algorithm was experimented using WEKA software. All the classifiers were analyzed and the most successful algorithms were selected (Bagging, IBK, Random committee, random forest, Bayes, Function, Lazy, MA meta, Misc, Rules and Trees), no any result was generated for M1 because it does not support the file format (.arff) and as such it didn't run.

The study has contributed in providing alternative to diagnose breast cancer at early stage by avoiding other methods such as X-ray, mammography which subject patients to high radiation. However, referring to the dataset there is no significant disparity, discrepancy or difference between the number of positive and negative labels. Similarly, treatment options are recommended based on the stage of the cancer which the study is not able to take care of.

The work of Saria & Huda (2019), where a review was presented on breast cancer diagnosis and prediction using both machine learning and data mining technique. The study aimed at reviewing te role of machine learning and data mining technique in breast cancer detection and diagnosis. The study first outlined the types of breast cancer, te variations in treatment been administered. It also ignited the relationship of data mining and machine learning, thereby explaining the pros and cons of different data mining algorithms.

Twenty four research articles were reviewed to ascertain the state of art and explore the computational methods to predict breast cancer. Most of the studies were comparing different classification techniques on a dataset to correctly classify if the tumor presented is cancerous (malignant) or non cancerous (benign). Therefore, an emphasis should be made in studying data sets of dedicated case studies. Thus, good data set that has minimal missing values provide better accuracies and in turn will help in developing improved predicting models with high impact and practical value.

4.0 Conclusion:

The study as reviewed some previous literatures to ascertain the state of art in revealing the detailed application of data mining techniques in health care, especially in predicting the presence of breast cancer. Future study will widen the horizon in analysing the most frequent data mining technique used in diagnosis of carcinoma cancer that starts in epithelial cells and spreads throughout the body that emerges from cells originating in the ectodermal germ layers.

References:

1. Agrawal, L., & Gupta, H. (2013). Optimization of C4.5 Decision Tree Algorithm for Data Mining Application. International Journal of Engineering Technology and Advanced Engineering, 1532 -1539.
2. Ahmed, A. S., Laila, A. A.-E., Sherif, K., & Ayman, A. G. (2018). Stage-Specific predictie models for main prognosis measures of breast cancer. Future Computing and Informmatics Journal, 3, 391-397.
3. Arafat, H., Barakat, S., & Goweda, F. (2012). Using Intelligent Techniques for Cancer Classification. . International Journal of Emerging Trends and technology in Computer Science, 1(3), 26-36.
4. C., N., & D., M. (2018). Breast cancer prediction system using Data mining methods. International Journal of Pure and Applied Matematics, 119(11), 10901-10911.
5. Cela, E., & Frasheri, N. (2013). Data Mining Techniques and Tools Used in Healthcare Databases. International Conference on Research and Education Challenges Towards the Future, 1-11.
6. Durairaj, M., & Ranjani, V. (2013). Data Mining Applications in Health Care Sector: A Study. International Journal of Scientifc and Tecnology Research, 2(10), 29-35.
7. Gupta, S., K. D., & Sharma, A. (2011).). Data Mining classification Techniques Applied for Breast Cancer Diagnosis and Prognosis. Indian Journal of Computer science and Engineering,, 2(2), 188-195.
8. Haifeng, W., & Sang Won, Y. (2015). Breast Cancer Prediction Using Data Mining Method. Proceedings of the 2015 Industrial and Systems Engineering Research Conference. S. Cetinkaya and J. K. Ryan, eds.
9. Hamza, S., & Nagendra, N. (2020). Data Mining Techniques in Predicting Breast Cancer. Journal of Applied Science, 20(4), 124-133.
10. Iffat, K., Ankita, G., Nandan, P., & Bindu, G. (2020). Breast Cancer Prediction Using Data MiningBreast Cancer Prediction Using Data Mining. International Journal of Scientific Research and Engineering Developmment, 3(2), 978-980.
11. Kharya, S. (2012). Using Data Mining Techniques For Diagnosis and Prognosis of cancer Disease. International Journal of Computer Science, Engineering and Information Technology, 2(2), 55-66.
12. Luo, Y., Houcberg, E., & Uzoner, O. (2015). Subgrapgh Augmented Non-Negative Tensor Factorization (SANTF) For Modelling Clinical Narrative Text. Journal of the American Medical Informatics Association, 1(13), 136-142.
13. Mumine, K. K. (2019). Breast Cancer Predioction and Detection Using Data Mining Classification Algorithm: A Comparative Study. Tehnicki Vjesnik, 26(1), 149-155. doi:https://doi.org/10.17559/TV-20180417102943

14. N., P., & R. Jagadeesh, K. (2016). Knowledge Discovery Stages for Early Detection of Harmful Cancer Disease. International Journal of Control Theory and Applications, 9(51), 196-202.
15. Padapriya, B., & T.Velmarugan. (2014). A Survey on Breast Cancer Using Data Mining Techniques. IEEE International Conference on Computational Intelliggence and Computing Research, 1234-1237.
16. R., P., & S. Vinila, J. (2019). A Research on Breast Cancer Prediction using Data Mining Techniques. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8(11S2), 362-370. doi:10.35940/ijitee.K1058.09811S219
17. Rozita, J. O., Nasroalla, M. K., & Saeid, A. a. (2017). Data mining and medical world: breast cancers' diagnosis, treatment, prognosis and challenges. American Journal of Cancer Research, 7(3), 610-627.
18. Saria, E., & Huda, K. (2019, April). Breast Cancer Diagnosis and Prediction Using Macone Learning and Data Mining Tecniique: A Revview. IOSR Journal of Dental and Medical Sciences, 18(4), 85-94.
19. Sowmiya, T., Gopi, M., Begin, L., & Robinson, T. (2014). Optimization of Lung cancer Using Modern Data Mining Technique. International Journal of Engineering research, 3(5), 309-314.
1. Vijaya, G., Suhasini, A., & Priya, R. (2014). Automatic Detection of Lung Cancer in CT Images. International Journal of Researc in Engineering and Tecnology, 3(7), 182-186.