ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

Role of Big Data Analytics in HealthCare System: A Systematic Review for Breast Cancer

[1] Simardeep Kaur, [2] Maninder Singh
Department of Computer Science, Punjabi University, Patiala

Abstract: **- A major area of research in big data exists in healthcare. As the amount of the data in Healthcare is growing at a rapid speed, there is a possibility of knowledge gain from analytics of such tremendous amounts of data. Big Data analytics in healthcare can improve patient care and also medical practice. The present study is an attempt to focus on systematic review of literature that how different big data methodologies are helpful for diagnosis and prognosis of the breast cancer patients. The findings showed that researchers observed a positive effect of data analytics in case of cancer patients. However, an issue related to technology acceptance exists in use of big data analytics in medical field. We will also examine possible future work to provide an approach to gain most of the knowledge in health informatics by highlighting some prominent areas of investigation.**

***Keywords*- Health informatics Big Data Electronic Health Records.**

I. INTRODUCTION

Big Data is a term coined for the large and complex datasets or combination of datasets, such that traditional database and software techniques are inadequate to deal with them. Tremendous amount of data that comes from variety of sources and in a variety of formats at an alarming velocity, is exploding day by day. This massive volume of data comes from everywhere: sensors data, social media data, public web, transaction records, machines log data, cell phones GPS signals to name a few. The amount of data in all fields is growing exponentially. Only in the last two years, 90% of the world's data has been reported. Big data is everywhere and processing such huge amounts of complex and dynamic data can provide businesses real insights. From a practical point of view, big data has the potential to benefit business with advanced analytics methods that extract value from the data, better operations and intelligent decisions [15].Big data is not only about finding insights from complex and huge data, also aims to answer the questions that were unanswered yet. Big data can be described by four characteristics: Volume, Variety, Velocity and Veracity. Volume implies amount of the data. Variety refers to the diversified sources and types of data forms and formats such as structured, semi-structured and unstructured. Velocity is the measure of the speed at which data flows in from sources. Finally, veracity of data refers to the noise, abnormality and error-freeness of data.

A revolution in terms of big data is under way in health care. Electronic health records (EHR), machine generated/sensor data, health information exchanges, patient registries, portals, and genetic databases, public records and so on are the major sources of big-data in healthcare area. Although complexities exist in healthcare data, still potential and benefit in analytics of big data solutions exist in this realm [16]. The big data's role in medical field is to build better health profiles and predictive models to diagnose and treat patients in a better way. Big data in healthcare is used to improve medical practices such as predict epidemics, cure disease, lowering costs, improve quality of life and avoid preventable deaths. However recent trend is towards the digitization of the healthcare data from stored printed form [16]. Breast cancer is one of the leading cause of death among women all over the world. Over one million cases of breast cancer are recorded and 600,000 deaths occur annually due to this. The classification of breast cancer data can help in accurate diagnosis of the patients as it allows distinguishing benign tumors from the malignant ones. Patients can be provided with right cure at a right time. The aim of this paper was to review the existing literature on how big data tools and methodologies help to improve the classification of the breast cancer patient's data and for the better medical assistance. The review of the big data analytics for breast cancer patients will help to view the tools and methodologies for the diagnosis and prognosis of the cancer patients. The rest of the paper is structured as follows: In section 2 we review the existing evidence literature on the effectiveness of big data analytics for cancer patients. The next two sections 3 and 4, include the methodology and the results of this study respectively. Finally section 5 presents some discussions and conclusions.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

II. METHODOLOGY

The main aim of our study is to review the existing literature to identify the role of big data analytics for the detection of the breast cancer and to identify the empirical papers in this context. For the literature review the strategy used was: < breast cancer- related keywords> AND <big data -related keywords> AND ["healthcare analytics" OR "breast cancer classification" OR "WDBC and WPBC " "Data mining- breast cancer prediction" " breast cancer diagnosis and prognosis" "breast cancer diagnosis via data mining"].

3.1 Inclusion and Exclusion criteria

We have only included those studies that were focused on the effectiveness of the big data analytics in the healthcare field especially Breast cancer. This was done because of the lack of the proper study comparisons about the effectiveness of the different classification algorithms in context of breast cancer. Selected studies have documented the use of Wisconsin breast cancer datasets: Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC), Surveillances Epidemiology and End Result (SEER) data set. So the focus was laid on the following types of data:

Electronic Health Records: The data generated between a Physician and a patient during the period of treatment and care such as patient details, demographic details, clinical records, administrative records and history records, is of large scale.

Patient Reported Outcomes: Patient reported outcomes also termed as PROs provide information regarding the health of the patients using various outcomes such as functional status of parts of body, symptoms and other health related information.

Genomics and Imaging Data: The collection of genomic and imaging data helps in the primary and secondary care of the patients. Various data mining tools use genomic and imaging data to discover the various patterns and relationships found in the human body. So the Big data techniques help to identify the various patterns among the genomic and imaging data so that a proper care to the patient can be provided. For example, in case of breast cancer, the mammographic images of the patient can provide information about the presence of axillary lymph node. Lymph node status is highly useful for the prognosis purposes. We have excluded some types of studies in this review and the criteria for the exclusion was as following:

a) The research surveys conducted on the use of big data analytics in healthcare field

b) The papers that provide just opinion and not actually implemented.

c) Papers that don't make use of data analytics to diagnose and prognose breast cancer datasets.

The main focus of the study is to focus on the data analytics for breast cancer datasets.

3.2 Data Extraction

The inclusion criterion was on basis of full-text review after a review of titles and abstract. In Table1. each paper that was the part of the study is recorded with following information:

- Reference
- The type and the source of the Big data
- The aim of the effectiveness evaluation such as comparison of different data mining classifiers, evaluating performance of the classifiers, improving quality of class prediction, detection of probability of breast cancer using classifiers, application of feature selection and hybrid methodology for diagnosis and prognosis purposes.
- The methodology that was followed
- Findings of the effectiveness evaluation i.e. which classifier or the combination of the classifier is best suited for the breast cancer diagnosis and prognosis.

IV. RESULTS

The results showed the analysis of various classification techniques that supported the processing of the breast cancer data. In terms of breast cancer data analytics, Wisconsin data sets: WDBC and WPBC were widely used datasets. Wisconsin data sets were processed in ten out of twelve studies.

The findings showed that the researchers have reported a positive effect of the data analytics for the diagnosis and prognosis of the breast cancer data sets. Out of the 10 studies, two studies Lavanya et al. [3] and Sridevi et al. [5] have suggested a feature selection technique for the analysis of the breast cancer data set. Salama et al. [1] and Kharya et al. [14] have proposed a multi-classifier methodology i.e. fusion of the classification techniques is involved to achieve better accuracy. Senturk et. al. [12] have suggested an early prediction of the cancer using data of diagnosed patients. Chaurasia et. al. [7] have proposed a prediction method among the different classifiers.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

Table1. Effectiveness of data analytics for breast cancer detection

Reference	Source of Big Data	Aim of Effectiveness evaluation	Methodology	Findings of the effectiveness evaluation
Purnami *et.* al. [1]	Wisconsin Data set with 699 samples	use of 1-norm SVM for feature selection and Smooth SVM (SSVM) for classification	Feature selection of breast cancer parameters, Breast cancer classification	Removal of weak parameters resulted increase in model accuracy.
Umesh D R *et.* al.[2]	Breast Cancer SEER Data set, 17 input variables	to accurately predict recurrence in dataset	Utilization of association rule mining	Resulted in a prediction model to predict breast cancer recurrence on SEER data set
Salama *et* al. [3]	Dataset / No. of Attributes / No. of instances WBC: 11 / 699 Wisconsin Diagnosis Breast Cancer: 32 / 569 Wisconsin Prognosis Breast Cancer: 34 / 198	Comparison among different classifiers, fusion of classifiers to identify most suitable multi-classifier approach	Use of chi-square test and Principal Component analysis (PCA) before application of classification algorithms	**WBC data set**: Integration of MLP and J48 with the PCA was reported superior to other classifiers. **WDBC dataset:**SMO only or integration of SMO and MLP or SMO and IBK is effective. **WPBC dataset** : Integration of MLP, J48, SMO and IBK is supassing.
Lavanya *et* al. [4]	Dataset / No. of Attributes / No. of instances Breast Cancer: 10 / 286 Breast Cancer Wisconsin (original): 11 / 699 Breast Cancer Wisconsin (Diagnostic): 32 / 569	Analyzing the potential of Decision tree classifier- CART with and without feature selection	Use of particular Feature selection using CART	Feature selection method varies from dataset to dataset. **Breast cancer dataset**: SVMAttributeEval method with accuracy 73.03 is best. **Breast cancer Wisconsin (original):**PrincipalComponents -AttributeEval method with accuracy 96.99% is best **Breast Cancer Wisconsin(Diagnostic) :**SymmetricUncertAttri-butesetEval with accuracy 94.72% is best.
Jojan*et* al. [5]	Dataset / No. of attributes / No. of instances SurviellanceEpdimiology and End Result (SEER): 17 / 215,950	Improving class prediction on imbalanced breast cancer dataset	Use of two-step approach 1) use of feature selection technique 2) use of over-sampling technique	C4.5 classifier achieved accuracy of 83.80%, sensitivity 85.17%, specificity 82.36%. C4.5 was better in classification than MLP and Naïve Bayes on SEER dataset.
Kharya *et* al. [6]	Wisconsin Datasets with 699 records with 9 medical attributes	Designing a GUI to input patients data and detection of probability of Breast Cancer using Naïve Bayes classifier	Performing prediction by mining patient's or data repository	The breast.D20.N699.C2.num dataset have 65.5% benign cases and 34.5 malignant cases. Naïve Bayesian classifier revealed to be efficient approach with maximum of 93% accuracy.
Chaurasia*et* al.[7]	UC Irvine machine learning repository: Breast-cancer-Wisconsin having 683instances (16 instances were removed with missing values) and 9 integer-valued attributes.	To develop accurate prediction model for breast cancer	Three algorithms: SMO, IBK and BF tree methods are Experimented in Weka toolkit	Among SMO, IBK and BF Tree methods, SMO results higher prediction accuracy i.e. 96.2% than IBK and BF Tree methods
Anagnosto-Poulos *et* al.[8]	Data set / No. of Attributes / No. of Instances	Handling breast cancer problem by employing two proposed neural network	To solve the diagnosis problem, a probabilistic approach is used.	For the diagnosis purpose neural classifier reaches 98% and for the prognosis problem it reaches 92%

	Wisconsin Diagnostic Breast Cancer (WDBC) 32 569 Wisconsin prognostic Breast Cancer (WPBC) 35 198	architectures over WDBC/WPBC datasets		
Sridevi *et* al. [9]	Data set / No. of Features / No. of Instances Wisconsin Diagnostic Breast Cancer (WDBC) 30 569 Wisconsin prognostic Breast Cancer (WPBC) 33 198	A feature selection algorithm Modified Correlation Rough Set Feature Selection (MCRSFS)	At level 1 feature selection on rough set and at level 2 feature selection on reduced set based on Correlation Feature Selection (CFS)	Resulted in reduced WDBC and WPBC data sets. Improves classification accuracy of almost all data mining algorithms. Multi-layer Perceptron (MLP) algorithm recorded 100 percent accuracy on reduced WDBC data set.
Senturk *et* al. [10]	Wisconsin data set includes 699 samples with 11 attributes	Early prediction of the breast cancer among patients under cover of data of the diagnosed patients	Use of RapidMiner 5.0 data mining tool	On basis of selected prediction model from the diagnosed patients, the diagnostics are predicted with some confidence values.
Kharya *et* al. [11]	Wisconsin Data set with 699 instances and 10 plus the class attribute. SEER dataset	Analysis of data mining approaches employed to diagnose and prognose breast cancer.	Review of journals and publications in medicine field.	Among soft computing approaches and data mining classifiers, decision tree is found to be best predictor with 93.62% accuracy on UCI and SEER dataset.
Kharya *et* al. [12]	Wisconsin Breast cancer Dataset (WBCD) taken from UCI repository with 699 instances and 10 + class attributes	To predict the status of the disease by employing a hybrid methodology	Evaluating two hybrid models for prediction: 1. Information Treatment and Option Extraction, 2. Decision Tree-Support Vector Machine	Decision tree- Support Vector Machine (DT + SVM) perform well for the classification of the breast cancer data. DT + SVM achieve accuracy 91 % with low error rate 2.58%

V. CONCLUSIONS

In this paper, we aimed to present a systematic overview of the literature to determine the role of big data applications that have helped healthcare to provide better patient care as well as better quality care. The results show that various researches have been focusing on the implementation of the various classification algorithms individually as well as in fusion to have benefits to diagnose and prognoses breast cancer patients. An early prediction would lead to better care and cure. We should like to highlight the fact that unpublished reports are not included in our survey.

Furthermore, we observed that the big data analytics in healthcare domain: breast cancer, presents a positive effect. It happens because of classification techniques that are applied to identify the various patterns and associations. Patients and clinicians can benefit from knowledge produced by big data analytics. The results shown are a step toward this objective. This survey was limited to a small subset of studies.

REFERENCES

1. Santi WulanPurnami, S.P. Rahayu and Abdullah Embong: Feature selection and classification of breast cancer diagnosis based on support vector machine, IEEE 2008.

2. Umesh D R and B Ramachandra: Association Rule Mining Based Predicting Breast Cancer recurrence on SEER Breast Cancer Data, IEEE 2015

3. Gouda I. Salama1 et. al.: Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers, International Journal of Computer Science and Information Technology (2277 – 0764) Volume 01- Issue 01,September 2012.

4. D. Lavanya: Analysis of feature selection with classification: Breast cancer datasets, Indian Journal of

Computer Science and Engineering (IJCSE), vol. 2, no. 5, pp. 756-763, October-November 2011.

5. Jojan J.: Duo Bundling Algorithms for Data Preprocessing: Case Study of Breast Cancer Prediction, Lecture notes on Software Engineering, Vol. 2, No. 4, November 2014.

6. S.Kharya: Naïve Bayes Classifier: A probabilistic Detection Model for Breast Cancer, International Journal of Computer Applications (0975-8887), Vol.92-No.10, April 2014.

7. Chaurasia V., Pal S.: A Novel Approach for Breast Cancer Detection using Data Mining Techniques, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 1, January 2014

8. Anagnostopoulous I., Anagnostopoulos C., Rouskas A., Kormentzas G., Vergados D.: The Wisconsin Breast Cancer Problem: Diagnosis and DFS time Prognosis using probabilistic and generalized regression neural classifiers, 2005.

9. T.Sridevi and A. Murgan: A Novel Feature Selection Method for Effective Breast Cancer Diagnosis and Prognosis, International Journal of Computer Applications (0975-8887), Vol. 88 – No.11, February 2014.

10. Senturk Z., Kara R.: Breast Cancer Diagnosis via Data Mining: Performance Analysis of Seven Different Algorithms, Computer Science & Engineering: An International Journal (CSEIJ), Vol. 4, No.1, February 2014.

11. S.kharya: Using data mining techniques for diagnosis and prognosis of cancer disease, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012.

12. S.kharya: Predictive Machine Learning Techniques for Breast Cancer Detection, International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 4(6), 1023-1028, 2013

13. Jimeng sun and Chandan K. Reddy: Big Data Analytics For Healthcare, SIAM international Conference on Data Mining, 2013.

14. Delen D., Walker G., Kadam A.: Predicting breast cancer survivability: comparison of three data mining methods, Artificial Intelligence in Medicine, vol. 34, pp. 113-127, 2005

15. VibhavariChavan and Rajesh N. Phursule: Survey Paper on Big Data, International Journal of Computer Science and Information Technologies, Vol. 5 (6), 2014, 7932-7939.

16. Ashwin Belle, RaghuramThiagarajan, S.M. Reza Soroushmehr, FatehmehNavidi, Daniel A. Beard and KayvanNajarian: Big Data Analytics in Healthcare, Hindawi Publishing Corporation BioMed Research International Volume 2015.

17. Thakur P.: Review Paper on Different Methodology for Cancer Detection, International Journal of Information and Technology (IJIT) - Vol. 2 Issue 3, May-Jun 2015.