Original Article
Millennium Journal of Health
Open access Journal

Prediction of malignant and benign tumors based on diagnosed clinical data: A case study, St. Paul's Hospital Millennium Medical College

Belay Alemayehu Debessa*

Affiliations

Department of ICT, St. Paul's Hospital Millennium Medical College, Addis Ababa, Ethiopia

Correspondence *

Belay Alemayehu

belay.alemayehu@sphmmc.edu.et

St. Paul's Hospital Millennium Medical College

Publication information

Received: 15-Oct-2022

Accepted: 29-Jun-2023

Published: 01-Jul 2023

Citation: Alemayehu B., Prediction of Malignant and Benign Tumor based on diagnosed clinical data case study St. Paul's hospital millennium medical college. MJH, 2023, Volume 2 (2): eISSN: 2790-1378.

Abstract

Background: Malignant growth is an exceptional human test. These days, healthcare prediction is a data analytics method focused on reducing future medical costs. The predictive technique uses a patient's medical history to evaluate all the potential health risks and predict future medical treatment in advance

Objectives: This study aimed to design a data analytics model that predicts the occurrence of cancer cells from St. Paul's Hospital Millennium Medical College medical data.

Methods: Prediction of malignant and benign tumors from the big medical data that has been collected by different academic and medical imaging departments at the St Paul's hospital millennium medical college is designed. Novel data engineering techniques are applied to ensure the quality of data and integrate data from different sources. A deep learning approach based on a logistic regression function is employed to build the model.

Results: The deep learning is implemented on a Hadoop framework by configuring five commodity machines, each of them consisting of a core i3 processor, 4 Giga Byte RAM, and 1 Tera Byte of hard disk storage. A classification system did classifications. The performance of such systems is commonly evaluated using the data in the confusion matrix. The prediction probability is almost 0.99. This is one of the most accurate probabilities.

Conclusion: This study introduced an approach to identifying cancer cell presence in patients. It provides a very appropriate basis to use promising software platforms for developing applications that can handle big data in medicine and healthcare.

Keywords: Data analytic; deep learning; medical history; medical image; model; predict cancer

Background

Cancer is one of the leading causes of death worldwide, killing nearly 10 million people in 2020, or nearly 1 in 6(1). Accurate assessment and classification of disease, especially cancer, which is of great importance in medicine, is still poorly understood and treatment planning is often based on trial and error (2). It is also complex and treatment outcomes vary widely from patient to patient. In recent years, a vast amount of data on cancer diagnosis and treatment has emerged due to the development of biomedical technologies and approaches. The potential of big data in healthcare opens new windows for improving clinical diagnosis or treatment, but there are many challenges in efficiently analyzing and interpreting such large and complex data. For example, managing, extracting, analyzing, integrating, visualizing, and communicating hidden information from the myriad data representations of cancer has become one of the greatest challenges in the next generation of biomedicine.

Data analytics is fundamentally changing methodologies, procedures, frameworks, and technologies traditionally used in detecting the occurrence of cancer cells. Thus, in this research, an attempt is done to predict future tumors as cancerous or non-cancerous. That can be possible by analyzing the different types of data collected on diagnosed cancer tissues.

In the systematic review by Mitchell et al. (3) healthcare provider delay related to initial misdiagnosis and insufficient examination by the practitioner, was the most commonly occurring theme associated with delay in referral relates to the study approach and research hypothesis in this study because it examines the factors contributing to provider or practitioner delay include: symptom misattribution, no examination or investigation of malignancy, co-morbidity, patient characteristics. It is pertinent to this study because the predisposing factors and enabling resources may contribute to a late-stage appraisal or treatment of cancer diagnosis.

Predictive analytics supports healthcare sectors to achieve a high level of effective overall care and preventive care, as predictive systems' results allow treatments and actions to be taken when all the risks are recognized in the early stages, which aids in minimizing costs. (4).

Health prediction is a data analysis method that focuses on future medical cost reduction. Predictive technologies use a patient's medical history to assess potential health risks and predict future treatment in advance (5). Loginov et al. (6) found that by retrieving and confirming previous patient data, information, and diagnoses from a database, predictive methods can be implemented through prediction, saving time and cost. Parkland Hospital in Dallas, Texas implemented a predictive system that scans all patient data and information to identify potential risks and consequences. As a result, the hospital saved more than half a million dollars. It has been used to monitor patients and prevent future complications, especially in predicting heart failure and illness (7).

This study aims to develop a predictive model that can accurately classify future tumors as either non-cancerous or cancerous. The ultimate goal is to provide valuable insights into cancer incidence and mortality in Ethiopia. Detecting cancer at an early stage is crucial, and to achieve this, it is essential to gain a precise understanding of the existing barriers to and delays in cancer care. Once known, effective prediction of cancer can be prioritized and resources allocated in a cost-sensitive manner. That can be possible by analyzing the different types of data collected on those diseases (8). The successful application of data analytics should be used to facilitate health planning and improve timely diagnosis and access to treatment, framed within the context of comprehensive cancer control and preventing death. It positively impacts people's lives through preventive medical strategies and individualized patient treatment. It also has developed and validated a prediction model to identify patients at high risk of cancers for prevention or further assessment. The model could be used to identify cancer cell presence in patients.

Methods

Study setting, design, period, and population

This study is qualitative, designing a data analytics model, that predicts the occurrence of cancer cells from diagnosed clinical data, which is medical images, biomedical signals, handwritten prescriptions, and structured data of the pathologist often it is assigned a pathological grade to a tumor according to how malignant the tissue looks under the microscope.

Data collection and tools

The data were collected cross-sectionally from 2010 to 2016 using the records of patients with malignant and benign tumors at St Paul's Hospital millennium medical college.

Data Processing and Analysis

The Hadoop framework for classifying cancer as either malignant or benign based on pathologically-proven diagnostic data as well as Logistic regression with stochastic gradient descent (SGD) algorithm is used in the proposed framework to develop the best prediction model. Logistic

regression is trained using the prior clinical records of the patients.

Classification algorithms like support vector machines (SVM), k-nearest-neighbor (KNN), decision trees (DTREE), or Bayes' classifier (BC) are well understood and widely applied. One might ask why my motivation for exploring LR as a fast classifier to be used in data mining applications is its maturity. LR is already well-understood and widely known. It has a statistical foundation that, in the right circumstances, could be used to extend classification results into a deeper analysis. Thus, the Big Data management challenge becomes one of being able to extract the required design points; the modeling problem reduces to a designed analysis with reduced noise and less potential for spurious correlations and patterns relative to a randomly selected sub-sample of the same size.

People think LR is slow, but that's not true. Approximate numerical methods, regularization, and efficient implementation make LR fast and better than modern algorithms. This new implementation uses a modified iteratively re-weighted least-squares estimation procedure. It can compute model parameters for sparse binary datasets with hundreds of thousands of rows and attributes, and millions or tens of millions of nonzero elements in just a few seconds.

Results

Socio-demographic characteristics of participants

A Hadoop cluster was set up using OpenStack on an underlying server whose specification is shown in Table 4.1. Five virtual machines were configured on the server. The hardware and software configuration for each of the virtual machines is shown in Table 4.2. Each virtual machine is assigned 1vcpu cone i3, 4GB RAM, and 1 TB of hard disk storage.

Table 4.1 Hardware specification

Hardware	CPU model	intel(R) core (TM) i3-3110M CPU @ 2.40GHz 2.40GHz
	Core	Corei3
	Hard disk	5TB
	Memory	25GB

The Hadoop-2.7.4 was used with a single vector machine (VM) configured as the NameNode and the remaining four VMs as DataNodes. The NameNode was not used as a DataNode. The replication level of each data block was set to 3. Two typical Hadoop MapReduce applications were run as Hadoop Yet another resource (YARN) jobs. The TeraGen application available as part of the Hadoop distribution was used to generate different sizes of input data.

Table 4.2 Software and hardware configuration of each Vector machine

Software	Operating System	Ubuntu 14.04.3 LTS
	JDK	Open Jdk 1.7
	Hadoop	2.7.2
	OpenStack	Nova
Hard Ware	CPU	1vCPUs
	Processor	Intel xeon
	Hard Disk	20GB
	Memory	2GB

To test the Mahout installation, execute the command: mahout This will list the available programs within the distribution bundle, as shown in the following Figure 4.1.

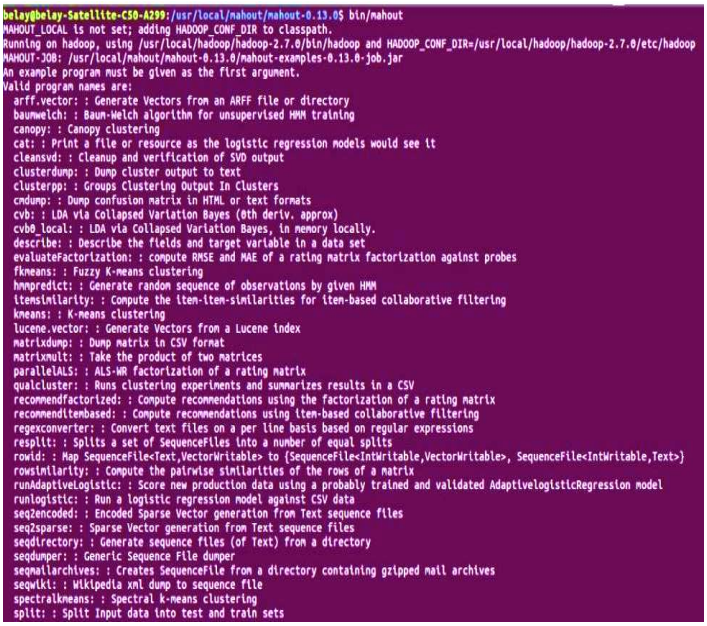

Figure 1. Screenshot of the mahout

This cancer dataset was obtained from St. Paul's Hospital Millennium Medical College where the samples arrive periodically. The database, therefore, reflects this chronological grouping of the dot. In this database, the following attributes exist Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitoses. The values of the attributes are between 1 and 10. Each instance of the dataset has one of two possible classes: non-cancerous indexed with 0 or malignant index with 1. The class distribution is for non-cancerous: 10656 (65.5%) and for Malignant: 5736 (34.5%).

Apache Mahout is a library of scalable machine-learning algorithms. Apache Mahout is implemented on top of Apache Hadoop and uses the MapReduce paradigm. Machine learning is a type of artificial intelligence focused on enabling machines to learn without being explicitly programmed, and it is commonly used to improve future performance based on previous outcomes. Big data is stored on the Hadoop Distribution File System (HDFS). Apache Mahout (2013) is

used to execute machine learning algorithms that extract meaningful patterns from datasets. Mahout's implementation of logistic regression using SGD supports the following command lines:

Training the model

bin/mahout trainlogistic --passes 100 --rate 50 --lambda 0.05 --input /usr/local/mahout/mahout-0.13.0/cancer/cancer22.csv --features 9 --output /usr/local/mahout/mahout-0.13.0/cancer/model --target Class --categories 2 --predictors Clump_Thickness Cell_Size_Uniformity Cell_Shape_Uniformity Marginal_Adhesion Single_Epi_Cell_Size Bare_Nuclei Bland_Chromatin Normal_Nucleoli Mitoses --types numeric

The outcome of the execution of the trainlogistic method is shown in the following Figure 4.2 below.

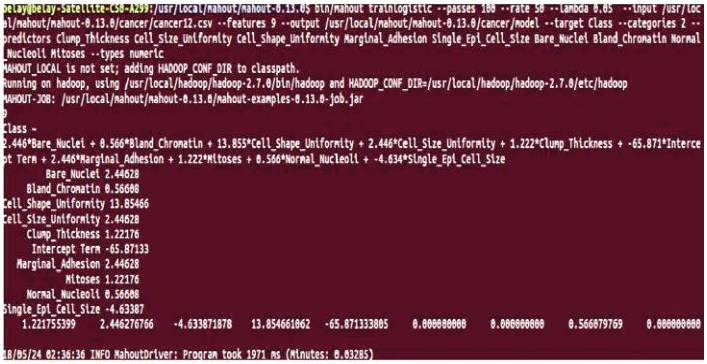

Figure 2 Training model screenshot

The important parameters for the trainlogistic function are explained in the following table

Table 4.3 Parameter description of the trainlogistic function

Parameter Name	Description
Input	This is the input dataset
Output	The model is saved as the name given
Target	This is the target variable field
Categories	This refers to the number of categories
Predictors	These are the predictor variable fields
Types	This is the list of types of the predictor variables
Features	This is the number of features

Passes: This specifies the number of times the input data should be re-examined during training. Small input files may need to be examined dozens of times. Very large input files probably don't even need to be completely examined.

Rate: This sets the initial learning rate. This can be large if you have lots of data or use lots of passes because it decreases progressively as data is examined.

Testing and evaluation

Now, let's evaluate the model generated using the dataset, using the following command:

bin/mahout runlogistic --input /usr/local/mahout/mahout-0.13.0/cancer/cancer22.csv --model /usr/local/mahout/mahout-0.13.0/cancer/model --auc --scores –confusion

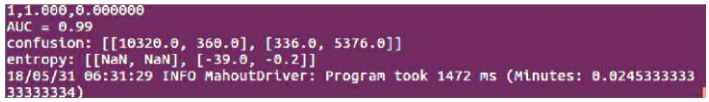

Figure 3. AUC screenshot

There are several methods to access the accuracy of the model; Among which the most widely used are the confusion matrix and the area under the curve.

The confusion matrix

A confusion matrix (9) contains information about actual and predicted. Classifications are done by a classification system. The performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two-class classifier.

The entries in the confusion matrix have the following meaning in the context of the study:

10320 is the number of correct predictions that an instance is benign.

360 is the number of incorrect predictions that an instance is cancerous.

336 is the number of incorrect predictions that an instance is benign, and

5376 is the number of correct predictions that an instance is cancerous. The confusion matrix (9) is shown in the following table

Table 4.4 Confusion matrix

		Predicted	
		cancerous	Benign
Actual	benign	10320	360
	cancerous	336	5376

The area under the curve

Accuracy is measured by the area under the Receiver Operating Characteristic (ROC) curve measure.

Discussion

A perfect model will achieve a true positive rate of 1 and a false positive rate of 0. A perfect model will score an Area Under the Curve (AUC) of 1, while random guessing will score an AUC of around 0.5. In practice, all models will fit somewhere in between. Now, these matrices show that the model is Having 0.99 as the value for AUC is good, but we will check this on test data as well.

The confusion matrix informs us that out of 10656 benign tumors, it has been correctly classified in 10320 instances and that 360 cancerous tumors are also classified as benign. In the case of cancerous tumors, out of 5736, it has been correctly classified 5376. This program makes accurate predictions. Remarkably, the prediction probability is almost exactly 1, even though any value of 0.5 or greater would be considered cancerous. The probability is still very high. This is one of the most accurate probabilities that could be found. This means that it has been predicted inaccurately for 1% of our training set.

The model could be used to identify cancer cell presence in patients. It provides a very appropriate basis to use promising software platforms for the development of applications that can handle big data in medicine and healthcare. One such platform is the open-source distributed data processing platform Apache Hadoop MapReduce which uses massive parallel processing (MPP) (10).

Delen et al (11), in their work, have created models for predicting the survivability of analyzed cases utilizing the SEER breast cancer dataset. Two algorithms, an artificial neural network (ANN) and a C5.0 decision tree, were employed to create prediction models. C5.0 gave an accuracy of 93.6% while ANN gave an accuracy of 91.2%. Logistic regression with SGD algorithm is used in the proposed framework to develop the best prediction model efficiently classifies the cancer disease with the accuracy of 99%

Once known, effective predictors of cancer could be prioritized and resources allocated in a cost-sensitive manner. The successful application of data analytics should be used to facilitate health planning and improve timely diagnosis and access to treatment, framed within the context of comprehensive cancer control and preventing death.

The huge dataset is extensively generated in every industry sector. A physician is willing to extract useful information from the transactions to make the best decision; researchers are expecting to extract useful information from the experimental results and thus develop new theories and products; doctors need to extract useful information from the data model to determine the direction of disease. Thus, how to realize parallel data mining algorithms to improve the execution speed is becoming a significant problem. It requires efforts from all sectors to achieve the highest state of data mining.

In the future, the Hadoop platform should be further directed to improve its performance and efficiency. In logistic regression algorithms, the random partition method brings instability to experimental results; some reasonable methods should be developed to optimize logistic regression classification algorithms. Meanwhile, MapReduce programming can be further optimized, such as the big dataset can be compressed and the small dataset can be merged during the data transfer process. The parallel implementation of other clustering, classification, and association rules algorithms in Mahout should gain more attention in the future. Besides, Hadoop configuration parameters have a significant impact on the performance of Hadoop clusters; it can improve abilities for processing large-scale data by modifying Hadoop configuration parameters.

Conclusion

This paper proposes a classification model that deals with binary labels. The classifiers used for classifying this dataset and the various feature reduction techniques applied can be used for other classification problems, which involve categorizing the data into binary classes. Integrating various lexicons into this classification model makes this model classify the data that consists of categorical classes.

It is observed that the proposed prediction model efficiently classifies the cancer disease with an accuracy of 99% so it can be implemented on the free software Hadoop framework. Logistic regression is trained using the prior clinical records of the patients. To gain insight into how they can improve service while reducing costs, healthcare payers and providers are turning to data and analytics. Leading organizations are treating data as a strategic asset and putting processes and systems in place that help healthcare professionals improve decision-making and drive actionable results.

This work can be used for other areas related to classification problems by making some adjustments to the multi-tier predictive model and by using various context-specific lexicons.

Abbreviations

LR: Logistic regression

BC: Bayes' classifier

VM: Virtual Machine

YARN: Yet Another Resource

HDFS: Hadoop Distribution File System

AUC: Area Under the Curve

SPHMMC: Saint Paul's Hospital Millennium Medical College

Declarations

Consent for publication

Not applicable.

Ethical declaration

The study was approved by SPHMMC's Institutional Review Board (IRB). Permission was obtained from St. Paul's Hospital millennium medical college for cancer patients' data, and records of patients with malignant and benign tumors. The study was conducted based on the approved protocol following the Helsinki Declaration principles.

Acknowledgments

The authors acknowledge SPHMMC for funding the study and our special gratitude goes to the children and their families who participated in this study.

Funding

SPHMMC funds this study.

Competing interest

The authors read and approved the final manuscript. The authors declare that they have no competing interests.

Availability of data and materials

The datasets used in the current study or data collection tool are available from the corresponding author with a reasonable request.

References

1. Arbyn M, Weiderpass E, Bruni L, de Sanjosé S, Saraiya M, Ferlay J, Bray F. Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. The Lancet Global Health. 2020 Feb 1;8(2): e191-203.
2. Ferlay J, Partensky C, Bray F. More deaths from pancreatic cancer than breast cancer in the EU by 2017. Acta oncologica. 2016 Oct 2;55(9-10):1158-60.
3. Neal RD, Tharmanathan P, France B, Din NU, Cotton S, Fallon-Ferguson J, Hamilton W, Hendry A, Hendry M, Lewis R, Macleod U. Is increased time to diagnosis and treatment in symptomatic cancer associated with poorer outcomes? Systematic review. British journal of cancer. 2015 Mar;112(1): S92-107.
4. Conley TG, Hansen CB, Rossi PE. Plausibly exogenous. Review of Economics and Statistics. 2012 Feb 1;94(1):260-72.
5. Wheeler DC, Wang A. Assessment of residential history generation using a public-record database. International journal of environmental research and public health. 2015 Sep;12(9):11670-82.
6. Dmitriev AA, Kashuba VI, Haraldson K, Senchenko VN, Pavlova TV, Kudryavtseva AV, Anedchenko EA, Krasnov GS, Pronina IV, Loginov VI, Kondratieva TT. Genetic and epigenetic analysis of non-small cell lung cancer with NotI-microarrays. Epigenetics. 2012 May 1;7(5):502-13.
7. Arena R, Mathews CE, Kim AY, Lenz TE, Southern PM. Prevalence of antibody to Trypanosoma cruzi in Hispanic-surnamed patients seen at Parkland Health & Hospital System, Dallas, Texas. BMC research notes. 2011 Dec;4(1):1-4.
8. Zhang, B., Zhou, X., Zhu, C., Song, Y., Feng, F., Qiu, Y., ... & Wang, J. (2020). Immune phenotyping based on the neutrophil-to-lymphocyte ratio and IgG level predicts disease severity and outcome for patients with COVID-19. *Frontiers in molecular biosciences*, *7*, 157.
9. Kohavi R, Provost F. Confusion matrix. Machine learning. 1998 Feb;30(2-3):271-4.
10. Ristevski B, Chen M. Big data analytics in medicine and healthcare. Journal of integrative bioinformatics. 2018 May 10;15(3):20170030.
11. Delen D. Analysis of cancer data: a data mining approach. Expert Systems. 2009 Feb;26(1):100-12.
12. Komarek, P. (2004). *Logistic regression for data mining and high-dimensional classification*. Carnegie Mellon University.
13. Anil, R., Capan, G., Drost-Fromm, I., Dunning, T., Friedman, E., Grant, T., ... & Yilmazel, Ö. (2020). Apache mahout: machine learning on distributed dataflow systems. *The Journal of Machine Learning Research*, *21*(1), 4999-5004.