Frontiers in Health Informatics
2024; 13: 182
www.ijmi.ir
Open Access

Using machine learning algorithms in determining the stage of breast cancer from pathology reports

Shirin Samadzad-Qushchi[1], Parinaz Eskandarian[2], Zahra Niazkhani[3,4], Ali Rashidi[5], Habibollah Pirnejad[4,6]*

[1]Population Based Cancer Registry, Urmia University of Medical Sciences, Urmia, Iran
[2]Department of Computer Engineering, Azad University, Zanjan Branch, Zanjan, Iran
[3]Nephrology and Kidney Transplant Research Center, Urmia University of Medical Sciences, Urmia, Iran
[4]Erasmus School of Health Policy & Management, Erasmus University Rotterdam, Rotterdam, The Netherlands
[5]Department of Health Information Technology, Urmia University of Medical Sciences, Urmia, Iran
[6]Patient Safety Research Center, Clinical Research Institute, Urmia University of Medical Sciences, Urmia, Iran

Article Info

Article type:
Research

Article History:
Received: 2023-09-26
Accepted: 2023-12-13
Published: 2024-02-01

**** Corresponding author:***
Habibollah Pirnejad

Patient Safety Research Center, Clinical Research Institute, Urmia University of Medical Sciences, Urmia, Iran

Email: pirnejad@eshpm.eur.nl

Keywords:
Breast Cancer
Pathology Reports
Text Mining
NLP
TNM Stage
Machine Learning

A B S T R A C T

Introduction: After a cancer diagnosis, the most important thing is to determine the stage and grade of the cancer. Pathology reports are the main source for cancer staging, but they do not contain all the information needed for the staging. However, the text of these reports is sometimes the only available information. We were interested in knowing whether text mining methods can be used to predict staging only from pathology reports.

Material and Methods: A total of 698 pathology reports of breast cancer cases and their TNM staging collected from multiple centers in West Azerbaijan Province, Iran were used for this study. After preparing the semi-structured reports, the texts of the reports were imported into a program written by Python V3. Three machine learning algorithms of Logistic Regression, SVM, and Naïve Bayes and a simple pipeline were used for the purpose of text mining. The performance of the algorithms was evaluated in terms of accuracy, precision, recall, and F1 score.

Results: The Naïve Bayes algorithm achieved excellent results and a value rate of higher than 91% in all evaluation criteria (accuracy, precision, recall and F1 score). This means that the Naïve Bayes algorithm could classify the reports with high efficiency and its predictions were more correct than the other two algorithms. Naïve Bayes also outperformed SVM and Logistic Regression in terms of accuracy, recall and F1 score. In addition, Naïve-Bayes showed faster inference due to its simplicity and lower computational and training time.

Conclusion: We suggest using the proposed design in this study for predicting breast cancer staging, where there is a need but not all necessary information except pathology reports. This method may not be a useful for clinical management of cancer patients, but it can be safely used for epidemiological estimations.

Cite this paper as:
Samadzad-Qushchi S, Eskandarian P, Niazkhani Z, Rashidi A, Pirnejad H. Using machine learning algorithms in determining the stage of breast cancer from pathology reports. Front Health Inform. 2024; 13: 182. DOI: 10.30699/fhi.v13i0.519

INTRODUCTION

Today, cancer is one of the major public health problems in the world. After cardiovascular disease, cancer is the second leading cause of death in developed countries, and one in four deaths is related to cancer [1, 2]. Breast cancer is the most common cancer in women in most countries [1]. Most breast cancers are diagnosed in the late stage and between the ages of 15 and 49 [3], while early detection and diagnosis before the disease progresses is very important and increases the patient's survival [4]. If breast cancer is correctly diagnosed in its early stages, its progression can be largely prevented, and people can be saved from the risk of death. The five-year survival rate of breast cancer patients after early diagnosis is 88%, and the ten-year survival rate after early diagnosis is 80% [5]. Therefore, early diagnosis and stage estimation of breast cancer are among the important and great objectives of many researches.

[6].

In tumor staging, three general categories are distinguished/estimated: primary tumor size (T stage), number of involved lymph nodes (N stage), and metastasis (M stage) [7]. Staging of breast cancer is useful for clinicians to estimate the prognosis and subsequently the survival of patients [8]. It is also important in choosing appropriate treatment for cancer patients [9]. Based on the size of the primary tumors, the T-stage of a tumor can be: T0, T1, T2, T3, and T4 [10]. Based on the number of lymph nodes involved, the N stage of a tumor can be: N0, N1, N2 and N3 [7, 8, 10]. Metastasis is divided into two parts, M0 and M1, where M0 indicates no metastasis and M1 indicates involvement of other organs and parts of the body such as bone, brain, lung, and liver in cancer [7, 8, 10]. According to the above divisions, the stages of breast cancer are named from stage zero to stage four, and some of these stages also include subgroups that are indicated by the letters A, B, and C. As a rule, the lowest number indicates the lowest rate of cancer progression, and the highest number indicates the highest rate of cancer progression. Also, in the subgroups, the letter that comes first indicates the lowest (least dangerous) stage [11].

In addition to clinical pathology reports, many other sources of information are needed for this process. As a result, cancer staging is considered a difficult and time-consuming task. With the lack of necessary information recorded in patients' medical records, this important process was even considered impossible. Pathology reports, as the most accurate source of cancer diagnosis and staging, are most often available. These reports are written by clinical pathologists in the form of semi-structured text [12]. Given the importance of cancer staging and the existing problem of finding other sources of information from patients' medical records, we were interested to know whether the pathology reports of breast cancer can be used as the sole source for staging this cancer.

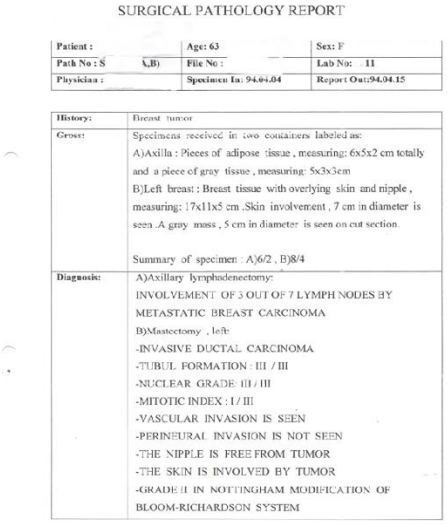
Text mining has become a broad approach for analyzing natural language textual data and extracting facts, knowledge, and patterns in a structured form from various textual data sources [13]. Text mining as a potential solution has been considered in many similar studies [14-16]. It can be done in different areas including rule-based approaches, natural language processing and machine learning [17]. In 2020, Pratiksha and colleagues conducted a study in India to extract breast cancer stage from pathology reports using natural language processing [15]. This study also used machine learning methods and a rule-based approach. The average accuracy was found to be 87%. In another study from three public hospitals including Massachusetts General Hospital, Birmingham Women's Hospital and Newton Wellesley Hospital, machine learning approach was used to detect 20 different categories of information from pathology reports of breast cancer with an average accuracy of 97% for all categories [14].

We know that the structure and content of clinical pathology reports change from one context to another. There are apparent variations in the way pathologists report common pathologic diagnoses. A study, for example, found 124 ways of saying invasive ductal carcinoma and 95 ways of saying invasive lobular carcinoma [18]. And to our knowledge, text mining has not been used in predicting cancer stage from pathology report in Iran. Thus, in this study, we were interested in knowing whether text mining methods can be used to predict staging from pathology reports.

MATERIAL AND METHODS

Data acquisition

The data were semi-structured texts of clinical/surgical pathology reports of breast cancer cases and their TNM staging. The reports were collected from the population-based cancer registration system of West Azerbaijan (WA) Province, Iran. We used the pathology reports of all 698 definite breast cancer cases that were recorded from multiple centers in WA province from 03/20/2014 to 03/20/2015. This dataset was used because it was the dataset with the most accurate TNM staging. An example of anonymized pathology reports is shown in Fig 1.



SURGICAL PATHOLOGY REPORT

Patient :	Age: 63	Sex: F
Path No : S (A,B)	File No :	Lab No: 11
Physician :	Specimen In: 94.04.04	Report Out:94.04.15

History:	Breast tumor
Gross:	Specimens received in two containers labeled as: A)Axilla : Pieces of adipose tissue , measuring: 6x5x2 cm totally and a piece of gray tissue , measuring: 5x3x3cm B)Left breast : Breast tissue with overlying skin and nipple , measuring: 17x11x5 cm .Skin involvement , 7 cm in diameter is seen .A gray mass , 5 cm in diameter is seen on cut section. Summary of specimen : A)6/2 , B)8/4
Diagnosis:	A)Axillary lymphadeneectomy: INVOLVEMENT OF 3 OUT OF 7 LYMPH NODES BY METASTATIC BREAST CARCINOMA B)Mastectomy , left -INVASIVE DUCTAL CARCINOMA -TUBUL FORMATION : III / III -NUCLEAR GRADE: III / III -MITOTIC INDEX : I / III -VASCULAR INVASION IS SEEN -PERINEURAL INVASION IS NOT SEEN -THE NIPPLE IS FREE FROM TUMOR -THE SKIN IS INVOLVED BY TUMOR -GRADE II IN NOTTINGHAM MODIFICATION OF BLOOM-RICHARDSON SYSTEM

Fig 1: A sample of breast cancer pathology report

Text mining pipeline design

We used Python V3 to develop the text mining pipeline. For this purpose, the modules/libraries

shown in Fig 2 were imported from Python V3.

```
import pandas as pd
    from nltk.corpus import stopwords
    from nltk.tokenize import word_tokenize
import string
    from gensim.models import Word2Vec
import numpy as np
    from sklearn.linear_model import LogisticRegression
    from sklearn.metrics import accuracy_score, precision_score, recall_score, F_score
    from sklearn.svm import SVC
    from sklearn.model_selection import train_test_split
    from sklearn.preprocessing import StandardScaler
    from sklearn import datasets
```

Fig 2: The required Python libraries for this study.

Data Preparation

The dataset had to be prepared so that the program could read it correctly. However, we were intended to perform our research with minimal data preparation. To prepare the data, we define a CSV file with two columns separated by commas. In the first line we write the names of the columns and there is no data in this line. From the second row to the end of the file, each row represented a separate data item containing the input and output of the machine learning model. We have divided the data set in the data matrix into four smaller matrices that are used in the training and evaluation phases of the machines.

The matrix X_train contained the inputs of the machine and the matrix y_train contained the outputs of the machine in the training phase. The matrix X_test contained the inputs of the machine and the matrix y_test contained the outputs of the machine in the evaluation phase. We determined that 80% of the input data set is used for the training phase and 20% of it is used in the evaluation phase.

Since the input text of the pathology report may contain additional and meaningless characters such as dots and parentheses, we had to perform an initial processing on the X_train and X_test matrices to remove these additional characters from the input text. It is discussed that for a good classification performance a feature selection method may not be necessarily because feature selection is not stable with small samples and high dimensionality [19]. Therefore, did not use feature selection methods in our study. Later, we used the Word2Vec tool from the gensim library to convert the texts in the two matrices of X_train and X_test into vectors of numbers. Word2Vec was preferred over the other word embedding methods because the training speed in Word2Vec is higher than similar models. Moreover, it overcomes the limitations of BoW and TF-IDF by preserving contextual information and representing words in a dense vector space [20].

To implement the proposed pipeline, a program was written in Python V3.

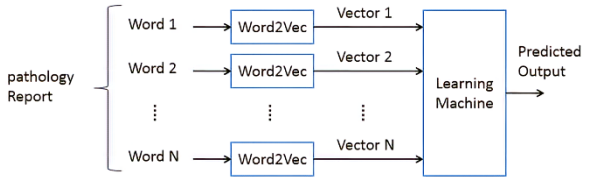


Fig 3: The overview of proposed pipeline.

At the beginning of the program, we needed to call a number of Python libraries (Fig 2) so that we could use the classes and functions defined in those libraries in the rest of the program. To achieve this goal, we needed a machine learning model that would take the text of the report as input and identify the key words (tokens) and the way they are combined in the pathology reports for each breast cancer stage. Then the machine learning model could provide the correct prediction. The training machine causes the internal parameters of the machine to be adjusted to make an appropriate prediction.

An overview of the proposed pipeline is shown in Fig 3. The words of the pathology report were input to our machine; they all entered the machine in parallel. But before entering the machine, they need to be converted into numbers. For this purpose, we use the Word2Vec tool. The function of the Word2Vec tool is to create a set of word vectors. When vectors are close together in vector space, their words have similar meanings based on context, and when word vectors are farther apart, their words have different meanings.

We used Naïve Bayes, Support Vector Machine (SVM), and Logistic Regression algorithms as machine learning models. The Naïve-Bayes algorithm is a machine learning algorithm derived from Bayes' Theory of Probability and used for classification problems. This algorithm is considered good for text classification where your training data set is high dimensional. This algorithm has been successfully used in sentiment analysis, new article classification, and spam filtering [21-23]. There are two models of Naïve-Bayes classification machine, which are known as Gaussian and Multimodal. In this research, we used the Gaussian model, which provides higher accuracy for the classification of our data set. SVM is one of the supervised learning methods used for linear data classification as well as regression problems [24]. Logistic regression is a machine learning classification algorithm that is also used to classify variables. In logistic regression, the dependent variable is a binary variable. It is a generalized linear regression method for learning the mapping from any number of numeric variables to a binary or probabilistic variable [14].

Evaluating performance of the text mining algorithms

We collected 698 pathology reports related to breast cancer patients. We used 80% of the selected datasets for training and 20% for evaluation. To evaluate the proposed scheme, we tested our three different machine learning models. The results of testing all the algorithms are visually presented and compared. Table 1 shows the results obtained by the studied algorithms respectively in terms of accuracy, precision, recall and F1 score criteria. In this study, in order to avoid being influenced by factors such as the training dataset, each of the studied algorithms was executed 25 times independently.

RESULTS

Table 1 shows the best and worst results obtained by testing the algorithms in 25 independent runs. The best and worst results obtained by the algorithms in their different implementations are shown in Fig 4. As can be seen, the Naïve Bayes algorithm has achieved excellent results, reaching a value of more than 91% in all evaluation criteria. This means that the Naïve Bayes algorithm was able to classify the pathology reports with high accuracy and precision. It also means that Naïve-Bayes algorithm can predict the stage of breast cancer more correctly (accuracy = 0.9196) than the other two algorithms. It performed better than SVM and better than logistic regression with 92.31% accuracy. Naïve Bayes showed faster inference due to its simplicity and lower computational and training time. The two evaluation criteria of recall and F1 score showed the very high performance of this algorithm compared to the other two algorithms.

Table 1: The best and worst values of 25 times testing the algorithms performance.

Indices	F1-score		Recall		Precision		Accuracy	
Values	The worst value	The best value	The worst value	The best value	The worst value	The best value	The worst value	The best value
Logistic Regression	0.293	0.336	0.463	0.502	0.214	0.252	0.463	0.502
SVM	0.282	0.367	0.453	0.529	0.205	0.28	0.453	0.529
Naïve-Bayes	0.9201	0.9201	0.9196	0.9196	0.9231	0.9231	0.9196	0.9196

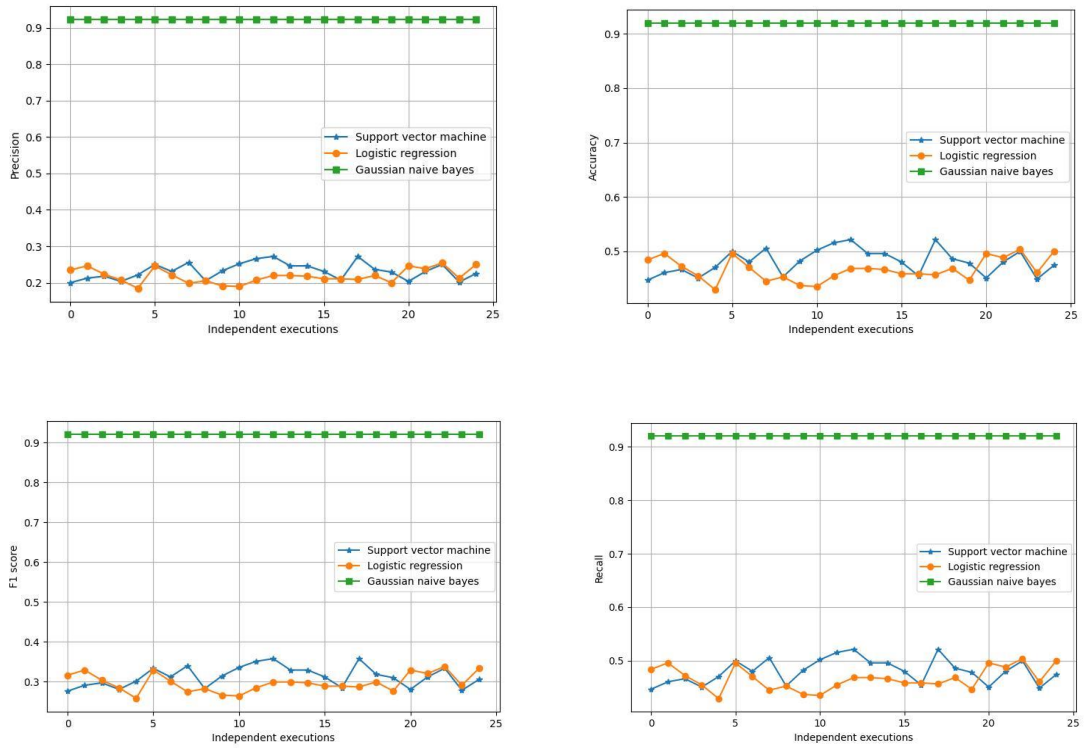


Fig 4: The performance indicators are shown for the text mining algorithms when they were run 25 times independently

It is obvious that the support vector machine algorithm did not achieve good results. In other words, this algorithm was only able to make 52%

correct diagnoses at best. In terms of prediction accuracy, it can be seen that this algorithm reached 27.2% in its best accuracy, which is not a satisfactory result. Like the support vector machine algorithm, the logistic regression algorithm did not achieve good results. According to the results obtained, it can be seen that the Logistic Regression algorithm has reached 50% accuracy in its best performance. This figure means that this algorithm was able to correctly predict only half of the data. As a result, it can be said that the logistic regression algorithm was not efficient enough for the purpose of this research.

DISCUSSION

TNM staging is crucial for effective treatment planning, determining disease prognosis, as well as epidemiological studies of cancer. In this study, machine learning techniques, specifically Logistic Regression, SVM and Naïve Bayes, were used to predict the TNM stage of breast cancer. Many studies have so far used machine learning algorithms for breast cancer classification and diagnosis [25]. Other studies tried to use different machine learning methods [26], combination of machine learning and rule-based approach [27], or very recently use large language models [28] to predict TNM stages of breast cancer from pathology text reports. The objective of the study was to propose a simple machine learning based model that can automatically (and with minimal data preparation) classify clinical/surgical pathology reports of breast cancer based on TNM stages.

In the proposed design, we considered a machine learning model that took the text of the pathology reports as input and identified the key words related to each stage of breast cancer. The evaluation showed that the Naïve Bayes algorithm achieved excellent results and a value rate of higher than 91% in all evaluation criteria (accuracy, precision, recall and F1 score). This means that the Naïve Bayes algorithm could classify the reports with high efficiency and its predictions were more correct than the other two algorithms. Naïve Bayes also outperformed SVM and Logistic Regression in terms of accuracy, recall and F1 score. In addition, Naïve-Bayes showed faster inference due to its simplicity and lower computational and training time. These findings about the performance of Naïve-Bayes are in line with some of the previous studies [29, 30].

The support vector machine algorithm did not perform well. In other words, this algorithm was only able to make 52% correct diagnoses at best. In terms of prediction accuracy, it can be seen that this algorithm has reached 27.2% in its best state, which is not a satisfactory result. The logistic regression algorithm has achieved 50% accuracy in its best performance. This figure means that this algorithm was only able to correctly predict half of the data. As a result, it can be said that the logistic regression algorithm is not efficient enough in the research.

Previous studies have reported high performance speed and time savings, more accurate predictions for multi-class problems, feature independence, and stronger performance even with a smaller number of training data for Naïve Bayes algorithm [31]. It also performs much better for categorical variables such as TNM stages than for other numerical variables. Furthermore, the evaluations performed on the proposed design showed that the larger the size of the training dataset compared to the test dataset, the higher the percentage of correct predictions [31, 32].

The current method of cancer staging is manual, which requires a high level of skill as well as a lot of time and resources [14]. The process is also prone to human error. With the increase of qualitative datasets and the advances in the application of machine learning algorithms, the ability to train on a large number of qualitative datasets is increased. Therefore, a better performance of the machines can be achieved and over time and this method can be expected to replace the manual method. An advantage of this method is that all necessary information for TNM staging is obtained from pathology reports and no additional source of information is necessary. Also, unlike manual methods, machine learning methods can make the prediction even when some important information, such as tumor size or number of involved lymph nodes, is missing from the pathology reports.

McCowan and colleagues [33] conducted a study similar to ours. They developed a classification system to determine the stage of lung cancer using machine learning techniques on pathology reports. However, they did not have a holdout set of data to test their system, and they could not predict the M stage of the cancer because metastasis was not often seen in pathology reports. Rajguru and colleagues [34] also developed a classification system based on Logistic Regression technique. For their study, in addition to pathology reports, they used various data sources, including ultrasound, mammography, radiology photographs, and so on. They achieved an accuracy of 95.90%. Our technique is more suitable for situations where supplementary data sources are not available due to poor recording system.

Our study had limitations that must be taken into account when interpreting the results. The dataset used to train the algorithms was large, although its dimensionality can be considered relatively high. The target values (i.e., TNM stages) used to train the algorithms in this study, although the best available, were the result of manual cancer staging and therefore prone to error. We didn't have the opportunity to test other powerful algorithms, such as deep neural networks, and this remains a topic for future research. We also do not know the

performance of our proposed design in predicting the staging of other types of cancer.

CONCLUSION

Our study showed that the proposed design using Naïve Bayes algorithm can generate its predictions correctly in an average rate of 91% of the cases. It produced more accurate predictions compared to SVM and logistic regression. Therefore, we suggest using the proposed design in this study for predicting breast cancer staging, where there is a need but not all necessary information except pathology reports. This may not be a useful method for clinical management, but it can be safely used for epidemiological estimations.

AUTHOR'S CONTRIBUTION

HP designed the study. SS collected the data. PS, HP, and SS prepared and analyzed the data and the paper's first draft. All other authors commented on analyzing data and drafting paper.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest regarding the publication of this study.

FINANCIAL DISCLOSURE

No financial interests related to the material of this manuscript have been declared.

ETHICS APPROVAL

This study was derived from the MSc. thesis of the first author under the supervision of the last author. The research ethics committee of Urmia University of Medical Sciences has reviewed and approved its proposal with the granting of ethical code of IR.UMSU.REC.1401.341.

REFERENCES

1. Ferlay J, Colombet M, Soerjomataram I, Mathers C, Parkin DM, Piñeros M, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. Int J Cancer. 2019; 144(8): 1941-53. PMID: 30350310 DOI: 10.1002/ijc.31937 [PubMed]
2. Khanjani N, Rastad H, Saber M, Khandani BK, Tavakkoli L. Causes of delay in seeking treatment in Iranian patients with breast cancer based on the health belief model (HBM). International Journal of Cancer Management. 2018; 11(6): e61383.
3. Norway CRo. Cancer in Norway 2018: Cancer incidence, mortality, survival and prevalence in Norway. The Cancer Registry of Norway Oslo; 2015.
4. Bray F, Grimsrud T, Haldorsen T, Johannesen T, Johansen A. Cancer in Norway 2008: Cancer incidence, mortality, survival and prevalence in Norway. The Cancer Registry of Norway; 2010.
5. Nguyen AN, Moore J, O'Dwyer J, Philpot S. Assessing the utility of automatic cancer registry notifications data extraction from free-text pathology reports. AMIA Annu Symp Proc. 2015; 2015: 953-62. PMID: 26958232 PMCID: PMC4765645 [PubMed]
6. Gaikwad SV, Chaugule A, Patil P. Text mining methods and techniques. International Journal of Computer Applications. 2014; 85(17): 42-5.
7. Weglarz G. Two worlds data: Unstructured and structured. Dm Review. 2004; 14: 19-23.
8. Baeza-Yates R, Ribeiro-Neto B. Modern information retrieval. ACM Press; 1999.
9. Moo TA, Sanford R, Dang C, Morrow M. Overview of breast cancer therapy. PET Clin. 2018; 13(3): 339-54. PMID: 30100074 DOI: 10.1016/j.cpet.2018.02.006 [PubMed]
10. Oskouei RJ, Kor NM, Maleki SA. Data mining and medical world: breast cancers' diagnosis, treatment, prognosis and challenges. Am J Cancer Res. 2017; 7(3): 610-27. PMID: 28401016 PMCID: PMC5385648 [PubMed]
11. Yan S, Qi Y. Apply text mining to advance cancer research. International Journal of Pharma Medicine and Biological Sciences. 2015; 4(2): 132-5.
12. Radha P, Preethi MBM. Text mining pathology and radiology records to habitually classify against disease: Computing The control of linking data sources. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 2018; 3(6): 76-84.
13. Sheikhpour R, Agha Sarram M, Zare Mirakabad MR, Sheikhpour R. Breast cancer detection using two-step reduction of features extracted from fine needle aspirate and data mining algorithms. Iranian Journal of Breast Diseases. 2015; 7(4): 43-51.
14. Yala A, Barzilay R, Salama L, Griffin M, Sollender G, Bardia A, et al. Using machine learning to parse breast pathology reports. Breast Cancer Res Treat. 2017; 161(2): 203-11. PMID: 27826755 DOI: 10.1007/s10549-016-4035-1 [PubMed]
15. Deshmukh PR, Phalnikar R. TNM cancer stage detection from unstructured pathology reports of breast cancer patients. International Conference on Computational Science and Applications. Springer; 2020.
16. Sufyan M, Shokat Z, Ashfaq UA. Artificial intelligence in cancer diagnosis and therapy: Current status and future perspective. Comput Biol Med. 2023; 165: 107356. PMID: 37688994 DOI: 10.1016/j.compbiomed.2023.107356 [PubMed]
17. Bhatia A, Victora CG, Beckfield J, Budukh A, Krieger N. "Registries are not only a tool for data collection, they are for action": Cancer registration and gaps in data for health equity in six population-based registries in

India. Int J Cancer. 2021; 148(9): 2171-83. PMID: 33186475 DOI: 10.1002/ijc.33391 [PubMed]

18. Buckley JM, Coopey SB, Sharko J, Polubriaginof F, Drohan B, Belli AK, et al. The feasibility of using natural language processing to extract clinical information from breast pathology reports. J Pathol Inform. 2012; 3: 23. PMID: 22934236 DOI: 10.4103/2153-3539.97788 [PubMed]

19. Kou G, Yang P, Peng Y, Xiao F, Chen Y, Alsaadi FE. Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. Applied Soft Computing. 2020; 86: 105836.

20. Asudani DS, Nagwani NK, Singh P. Impact of word embedding models on text analytics in deep learning environment: A review. Artif Intell Rev. 2023; 56: 10345–425. PMID: 36844886 DOI: 10.1007/s10462-023-10419-1 [PubMed]

21. Li C, Weng Y, Zhang Y, Wang B. A systematic review of application progress on machine learning-based natural language processing in breast cancer over the past 5 years. Diagnostics (Basel). 2023; 13(3): 537. PMID: 36766641 DOI: 10.3390/diagnostics13030537 [PubMed]

22. Zhang H, Li D. Naïve Bayes text classifier. International Conference on Granular Computing. IEEE; 2007.

23. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak. 2019; 19(1): 281. PMID: 31864346 DOI: 10.1186/s12911-019-1004-8 [PubMed]

24. Hearst M. What is text mining [Internet]. 2003 [cited: 15 Dec 2006]. Available from: https://people.ischool.berkeley.edu/~hearst/text-mining.html

25. Radak M, Lafta HY, Fallahi H. Machine learning and deep learning techniques for breast cancer diagnosis and classification: A comprehensive review of medical imaging studies. J Cancer Res Clin Oncol. 2023; 149(12): 10473-91. PMID: 37278831 DOI: 10.1007/s00432-023-04956-z [PubMed]

26. Bae JH, Han HW, Yang SY, Song G, Sa S, Chung GE, et al. Natural language processing for assessing quality indicators in free-text colonoscopy and pathology reports: Development and usability study. JMIR Med Inform. 2022; 10(4): e35257. PMID: 35436226 DOI: 10.2196/35257 [PubMed]

27. Kefeli J, Tatonetti N. Generalizable and automated classification of TNM stage from pathology reports with external validation. medRxiv. 2023; 2023: 2023.06.26.23291912. PMID: 37425701 DOI: 10.1101/2023.06.26.23291912 [PubMed]

28. Abedian S, Sholle ET, Adekkanattu PM, Cusick MM, Weiner SE, Shoag JE, et al. Automated extraction of tumor staging and diagnosis information from surgical pathology reports. JCO Clin Cancer Inform. 2021; 5: 1054-61. PMID: 34694896 DOI: 10.1200/CCI.21.00065 [PubMed]

29. Rathi M, Singh AK. Breast cancer prediction using Naïve Bayes classifier. International Journal of Information Technology & Systems. 2012; 1(2): 77-80.

30. Hazra A, Mandal SK, Gupta A. Study and analysis of breast cancer cell detection using Naïve Bayes, SVM and ensemble algorithms. International Journal of Computer Applications. 2016; 145(2): 39-45.

31. Harzevili NS, Alizadeh SH. Mixture of latent multinomial naive Bayes classifier. Applied Soft Computing. 2018; 69: 516-27.

32. Han-Joon K, Jiyun K, Jinseog K, Pureum L. Towards perfect text classification with Wikipedia-based semantic Naïve Bayes learning. Neurocomputing. 2018; 315: 128-34.

33. McCowan I, Moore D, Fry M. Classification of cancer stage from free-text histology reports. International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE; 2006.

34. Rajaguru H, Prabhakar SK. Expectation maximization based logistic regression for breast cancer classification. International Conference of Electronics, Communication and Aerospace Technology. IEEE; 2017.