Identification of Lung Cancer Metastasis from Patient Records: An Explanatory Meta-Diagnosis

Zeynep Ertem[1]*

1. System Science and Industrial Engineering Department State University of New York, Binghamton University, New York, United States of America.

Abstract

Even though survival rates are improving for many types of cancer due to improvements in cancer screening, treatment, and prevention, cancer is still the second-leading cause of death in the world. Knowing whether the disease is metastatic or not is crucial in the treatment of the disease. In recent years, many studies have reported on tools using machine learning in medicine with promising results in early prediction.

This paper studies a machine learning approach to identifying patients with metastatic cancer. Using basic supervised modeling approaches, this study shows the accuracy of classifying cancer conditions of patients a-posteriori into metastatic or non-metastatic.

Clinical article (J Int Dent Med Res 2022; 15(2): 955-958)

Keywords: Cancer, Metastasis, Gradient Boosting Trees, Logistic Regression, Machine Learning, Prediction, Forecasting, Supervised Modeling.

Received date: 16 March 2022 **Accept date:** 14 May 2022

Introduction

Cancer is one of the deadliest and costliest diseases both for the patient with cancer and for society. For example, it is estimated that more than 600K patients lost their lives in 2018, and the financial cost in the US is estimated to be more than $80 billion in 2015. Treatment plans are complex involving many decisions from healthcare professionals and patients often with uncertainty. Although several treatment options can usually be available, many ongoing efforts aim to develop new treatment options using clinical trials. While patients are assigned to groups in these clinical trials, the prediction of metastasis for cancer conditions of patients can be critical. Metastasis, the spread of malignant cells from a primary tumor to distant sites, poses the biggest problem to cancer treatment and is the main cause of death of cancer patients.

Several studies are published in early cancer diagnosis with machine learning approaches.[1,2,3] Using machine learning techniques like Naive Bayes (NB), Random Forest (RF), and Support Vector Machine (SVM) has been an emerging field to diagnose different cancer types by image analysis. Several studies implemented these machine learning approaches for the diagnosis of different cancer types like colorectal cancer[4,5,6,7,8], gastric cancer[9,10,11,12,13], brain cancer[14,15,16,17], breast cancer[18,19,20], rectal cancer[21] and lung cancer[22,23].

Accurate prediction of the risk of metastatic relapse can be critical for personalized treatment. Matching the patients to the right treatment plan is very related to the stage of the disease. In this work, my aim is to improve the matching process of the patients to clinical trials by using their patient records. Specifically, my goal is to find whether the patient is metastatic or not just by leveraging their medical records. In this paper, patient records are analyzed by basic machine learning methods to improve efficiency in clinical trials.1 Specifically, I use the healthcare records of lung cancer patients to classify patients' conditions into metastasis or non-metastasis.

Finding the stage of cancer in matching cancer patients to the clinical trials is important. Sometimes this information is readily available, but more often it is unavailable. One way to obtain this information can be to have medically trained professionals manually go through the

Corresponding author:
Zeynep Ertem
System Science and Industrial Engineering
Department State University of New York,
Binghamton University,
New York, United States of America
E-mail: zeynep@binghamton.edu

individual patient’s records. This manual approach is time-consuming and expensive. Our goal is to automate this process so that cancer status can be inferred from patient records to efficiently set up for the follow-up clinical trials (Figure 1).

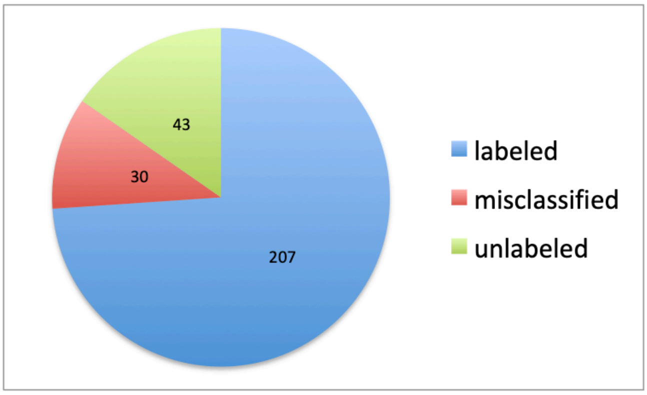

Figure 1. About 60% of the manually labeled cases are Metastasis.

Our approach is explanatory rather than predictive. Specifically, I use all patient records available a-posteriori to determine if the patient is metastatic or not instead of focusing on potential early indicators that might predict future metastasis. This is because our goal is to classify patients’ current statuses for future clinical trials as opposed to their future statuses.

Materials and methods

There are many known (and likely many unknown) indicators that correlate with metastatic cancer in de-identified patients’ records. For example, the therapies received, the lab results, genomic tests, appointment types, and frequencies can be used to classify if a patient’s cancer is in a metastatic stage. My data includes the status of 280 anonymized patients. Hence, I use basic supervised modeling approaches for this problem. I compare these supervised approaches to a baseline algorithm developed based on simple regular expression rules over patients’ records. Figure 1 shows the label distribution for the 280 manually labeled cases.

Data includes patient records that have information about therapies received, lab results, drugs used, genomic tests, appointment type and frequencies as well as healthcare professional notes in plain text. This data is used to create a feature set of size 18.

There are 280 manually labeled lung cancer patients. The baseline algorithm correctly classifies 207 of them, misclassifies 30, and leaves unclassified 43. In summary, the baseline algorithm is 73.9% accurate. Figure 2 summarizes the performance of the baseline algorithm on the manually labeled data set.

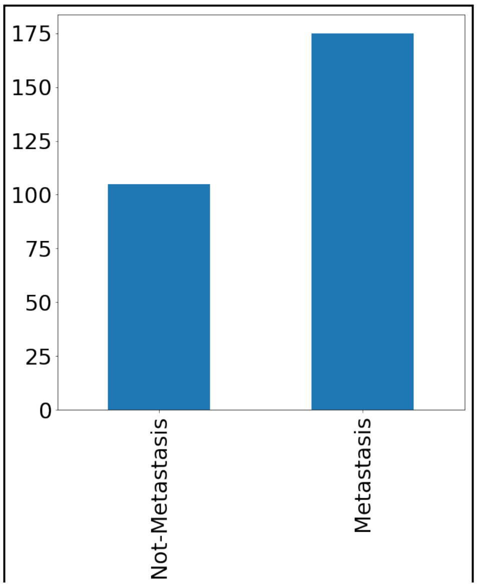

Figure 2. The baseline algorithm performance on the manually labeled dataset.

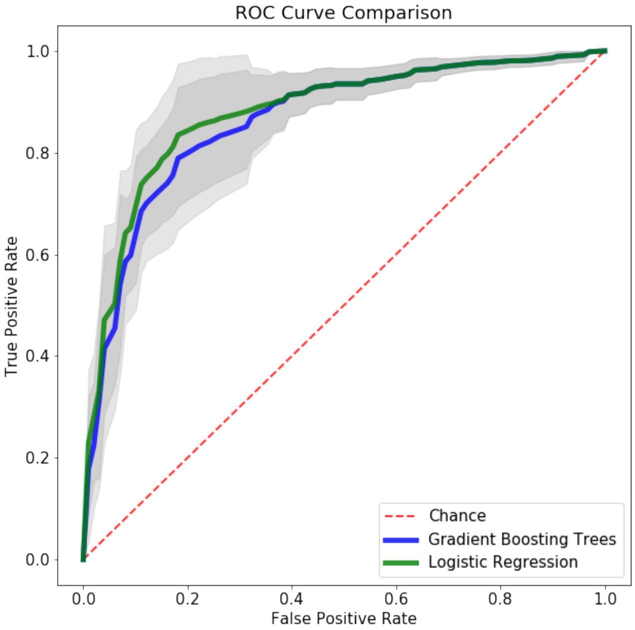

Figure 3. Accuracy comparison of supervised modeling approaches compared with the baseline algorithm.

I use two basic machine learning methods: (1) Gradient Boosting Tree Classifier

(GBT), and (2) Logistic Regression (LR). The reason I chose these models is two-fold. First, with these choices, I cover a basic non-parametric and a parametric model for this problem. Second, since I have a small amount of labeled data (only 280 instances), I use simple models over complicated models to avoid overfitting. Even GBT can get complicated depending on the settings of the classifier (e.g., number of trees, maximum depth), and I carefully set these parameters to avoid overfitting (Figure 3).

Results

Machine learning is a technique that employs a variety of statistical and optimization techniques that allows computers to learn from past examples. I studied two popular machine learning methods in this paper, Logistic regression, and Gradient Boosting Trees. It should be mentioned that these machine learning classification techniques can result in adequate and effective decision-making if the overfitting of the model is handled wisely.

To the best of our knowledge, this study is one of the early studies performed using medical health records data. Additional information about patient healthcare records can enhance the generalizability of the predictive model for their metastatic prediction.

Figure 3 compares the accuracy of supervised modeling approaches to the baseline algorithm. With supervised modeling approaches, the proposed approach can achieve an accuracy of about 88% with LR and 86% with GBT, whereas the baseline performance is at best 73%. The improvement with these algorithms is significant and these algorithms improve the accuracy of the classification of cancer conditions using healthcare records.

Furthermore, Figure 4 shows the corresponding ROC curves for LR and GBT obtained through 5-fold cross-validation. This study compares two mainstream classical machine learning methods. To implement these two machine learning algorithms, the data set is divided in the following manner: 80% is used in the training phase and 20% is used in the test phase. Prediction of lung cancer metastasis from the patient features data is an important step for the classification of the early metastasis stage. Decision-makers and healthcare providers can be alerted before the actual diagnosis of late-stage cancer.

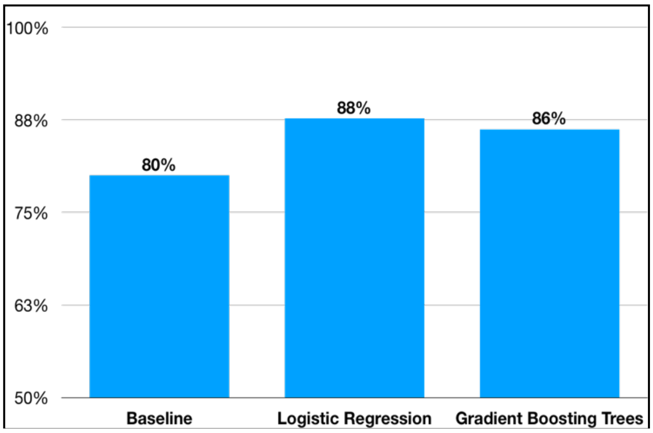

Figure 4. ROC curve obtained through 5-fold cross-validation for LR and GBT. Two approaches show similar performance.

Discussion

There are growing number of methods for prediction of the metastasis in different cancer types. To achieve accurate predictions, researchers are developing deep learning and artificial intelligence methods on the images. The increasing public availability of disease-related data is promising. However, using the patient records to predict the metastasis state is an area that hasn't been studied earlier.

In this study, we have used two main machine learning techniques to classify the patients with metastatic and nonmetastatic. This study will be a tool to alert the healthcare providers before they see a new patient or a current patient whether their features are alerting a metastatic state. This tool can also be extended to other cancer types. Using electronic health records in patient classification systems is novel and promising. Machine learning is a tool to utilize the creation and evaluation of algorithms that facilitate prediction, pattern recognition and classification. The classification is used to correctly place each observation in a category it belongs to.

Cross-Validation is a statistical technique; it is generally used to check and evaluate learning algorithms or models, by partitioning data into a learning set to train the model and testing set to evaluate it. The training set and the testing set in cross-validation are randomly divided into partitions (80% of data are in training sets and 20% of data are in testing sets). K-fold cross validation is used.

Accuracy by LR is comparable to GBT as can be seen form Figures 3 and 4. Both of these algorithms are fundamental modelling approaches to probability estimation in medial risk prediction.

While we believe that these insights are robust, they may reflect specific assumption. Mainly, superior performance of feature selection is likely biased by detailed description of the electronic health records. Thus, patient heath records should be accurate and complete for a clear picture.

This study was not without limitations. First, there was no independent study sample to perform external validation on these models. Also, the sample size may not have been sufficient for the data-hungry machine learning approaches.

Conclusions

In conclusion, this study revealed that the features from healthcare records are discriminative regarding the metastatic condition. Using basic supervised modeling approaches, my approach can improve the accuracy of classifying cancer patients a posteriori into metastatic or non-metastatic overusing simple rules based on regular expressions. Also, these models provide insights vis-á-vis the used data set. I find that, unsurprisingly, lack of mentions of metastasis in patient records is highly predictive of cancer status. Surprisingly, I find that zinc usage is also highly predictive of cancer status. This study was a single-center retrospective study. This study can be extended by collecting multi-center data for different cancer types to conduct a generalized application of machine learning based methods to classification of cancer conditions.

Declaration of Interest

The author reports no conflict of interest.

References

1. Bishop CM. Pattern recognition and machine learning. Springer, 2006.
2. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform 2007;11(2):59-77.
3. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2014;15(13):8–17.
4. Kainz P, Pfeiffer M, Urschler M. Segmentation and classification of colon glands with deep convolutional neural networks and total variation regularization. PeerJ 2017;3(5):3874.
5. Korbar B, Olofson AM, Miraflor AP, et al. Deep learning for classification of colorectal polyps on whole-slide images J Pathol Inform 2017;25(8):30.
6. Ichimasa K, Kudo SE, Mori Y, et al. Artificial intelligence may help in predicting the need for additional surgery after endoscopic resection of T1 colorectal cancer. Endoscopy 2018;50(3):230-40.
7. Bychkov D, Linder N, Turkki R, et al. Deep learning based tissue analysis predicts outcome in colorectal cancer Sci Rep 2018;8(1):3395.
8. Takamatsu M, Yamamoto N, Kawachi H, et al. Prediction of early colorectal cancer metastasis by machine learning using digital slide images. Comput Methods Programs Biomed 2019;178:155-61.
9. Lee JH, Choi MG, Min BH, et al. Predictive factors for lymph node metastasis in patients with poorly differentiated early gastric cancer. Br J Surg 2012;99(12):1688–92.
10. Kim HM, Pak KH, Chung MJ, et al. Early gastric cancer of signet ring cell carcinoma is more amenable to endoscopic treatment than is early gastric cancer of poorly differentiated tubular adenocarcinoma in select tumor conditions. Surg Endosc 2011;25(9):3087–93.
11. Kunisaki C, Takahashi M, Nagahori Y, et al. Risk factors for lymph node metastasis in histologically poorly differentiated type early gastric cancer. Endoscopy 2009;41(6):498–503.
12. Zhou CM, Wang Y, Ye HT, et al. Machine learning predicts lymph node metastasis of poorly differentiated-type intramucosal gastric cancer. Sci Rep 2021;11(1):1300.
13. Zhou C, Wang Y, Ji MH, et al. Predicting peritoneal metastasis of gastric cancer patients based on machine learning. Cancer Control 2020;27(1):1073274820968900.
14. Ambrosini RD, Wang P, O'Dell WG. Computer-aided detection of metastatic brain tumors using automated three-dimensional template matching. J Magn Reson Imaging 2010;31(1):85-93.
15. Farjam R, Parmar HA, Noll DC, et al. An approach for computer-aided detection of brain metastases in post-Gd T1-W MRI. Magn Reson Imaging 2012;30(6):824-36.
16. Pérez-Ramírez Ú, Arana E, Moratal D. Brain metastases detection on MR by means of three-dimensional tumor-appearance template matching. J Magn Reson Imaging 2016;44(3):642-52.
17. Sunwoo L, Kim YJ, Choi SH, et al. Computer-aided detection of brain metastasis on 3D MR imaging: Observer performance study. PLoS One 2017;12(6):0178265.
18. Yousefi S, Amrollahi F, Amgad M, et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. Sci Rep. 2017;7(1):11707.
19. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 2005;34(2):113-27.
20. Kim W, Kim KS, Lee JE, et al. Development of novel breast cancer recurrence prediction model using support vector machine. J Breast Cancer 2012;15(2):230-8.
21. Liang M, Cai Z, Zhang H, et al. Machine learning-based analysis of rectal cancer MRI radiomics for prediction of metachronous liver metastasis. Acad Radiol 2019;26(11):1495-1504.
22. Wang H, Zhou Z, Li Y, et al. Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from ^{18}F-FDG PET/CT images. EJNMMI Res 2017;7(1):11.
23. Wu Y, Liu J, Han C, et al. Preoperative prediction of lymph node metastasis in patients with early-T-stage non-small cell lung cancer by machine learning algorithms. Front Oncol 2020;10:743.