International Journal of Current Research and Review
DOI: http://dx.doi.org/10.31782/IJCRR.2021.131319

Scopus

Case Report

IJCRR
Section: Healthcare
ISI Impact Factor (2019-20): 1.628
IC Value (2019): 90.81
SJIF (2020) = 7.893

Application of Machine Learning for Improving Early Cancer Diagnosis

Jayasri Kotti

Professor, Computer Science and Engineering, LENDI Institute of Engineering and Technology, Vizianagaram-535005, AP, India..

ABSTRACT

Across the world, cancer becomes a catastrophe for a human being who is suffering from it. Cancer can be diagnosed at a premature stage to overcome the consequences at a later stage and the possibility of endurance considerably, as it can support appropriate medical action to patients. One of the frequently used innovative technologies for the diagnosis and detection of cancer is Machine learning (ML). In recent times ML has been used for the prediction and prognosis of cancer. Machine learning enables the creation of algorithms that can learn and make predictions. Various Machine Learning techniques can build a model to diagnose cancer based on finding accuracy level. It is possible for early detection of cancer through machine learning where we train the machine with previous data. This paper aims to predict cancer type based on symptoms given by the user. Here we adopted a supervised learning algorithm and then use the Logistic Regression based on accuracy and recall score i.e., the algorithm which gives high accuracy level and recall score. The proposed System executes with good performance as it generates accurate results.

Key Words: Machine Learning (ML), Data sets, Symptoms, Cancer, Logistic Regression, Supervised Learning

INTRODUCTION

Constant growth associated with cancer research has been achieved in the past few decades. For screening in the premature stage to find types of cancer before they cause symptoms different techniques came into existence. Researchers have been providing different innovative techniques and methods for cancer treatment. With the initiation of new techniques and methods in the field of medicine, a huge quantity of cancer disease data have been collected and are available to the medical study community. But the exact prediction of cancer is one of the remarkable and difficult tasks for doctors. For medical researchers, Machine Learning techniques and methods have become more popular. Machine Learning techniques can learn and recognize patterns and relationships between them from compound datasets, while they can successfully forecast future outcomes of a cancer disease. It is possible for early detection of cancer through machine learning where one can train the machine with previous data.

Nowadays Machine Learning techniques are being used in an extensive variety of applications ranging from identifying and classifying cancer via x-ray and CRT methods. According to the online statistics many articles have been published on the subject of Machine Learning and cancer disease. Still, the enormous majority of these papers are associated with using Machine Learning techniques to recognize, categorize, identify or discriminate cancer types and other tumours. The primary aim of cancer anticipates and prediction is different from the goals of cancer recognition and identification. Accomplishment in Machine Learning is not constantly assured. As with any technique, a good perceptive of the problem and approval of the restrictions of the data is important. Good quality of data is more important to get accurate results. The success rate in results occurs when we design and implement proper Machine Learning technique.

Machine Learning (ML) techniques repeatedly learn and improve with familiarity. Learning means recognizing and understanding the input data and making intelligent decisions based on the datasets. It is very composite to supply all the decisions based on all possible input dataset. To attempt these types of problems, algorithms are suggested. These algorithms construct information from exact data and past

Corresponding Author:
Jayasri Kotti, Professor, Computer Science and Engineering, LENDI Institute of Engineering and Technology, Vizianagaram-535005, AP, India.
Email: jayasrikotti@gmail.com; ORCID ID: 0000-0001-6501-1948

ISSN: 2231-2196 (Print) **ISSN:** 0975-5241 (Online)

Received: 09.11.2020 **Revised:** 03.01.2021 **Accepted:** 27.02.2021 **Published:** 05.07.2021

knowledge with the ideology of logic, probability and statistics. There are several ways to execute techniques in Machine Learning, and commonly used methods are supervised and non supervised learning. One of the Machine Learning techniques is classification. It uses known data to determine how the new data should be classified into a set of existing categories. A classification is a form of supervised learning. Figure 1 depicts the classification working process.

Literature Survey

In the world death rates are increased due to various types of cancers. Well, known types are lung cancer, breast cancer, blood cancer etc., and can be curable with early detection and treatment which varies from type to type.[1] Scientist has a pack of information such as text, facts and images which are properly separated that can be used by doctors to identify the type of disease.[2] Tumors can arise in any part of the body and can be transported to various other parts through blood flow in some cases. Early detection of its beginning stages could save a person's life.[3] million women every year are diagnosed with breast cancer, but most of them die due to late detection.[4, 10] Various methods are used for detection and prognosis of cancer diseases.[5] To discover hidden patterns and relationships advanced data mining techniques can be used.[6] For cancer progression Machine Learning techniques are very useful.[7] Artificial Intelligence has many branches which also includes Machine Learning that compiles various statistical probabilistic and optimization techniques that allow computers to learn from past datasets of various patterns.[8] Early detection of malignant stages reduces the risk of cancer spreading.[9] Many ML techniques are used to find the important risk factors. [11, 12] In medical sciences, ML techniques are very useful for solving prognostic and diagnostic problems. It is also useful in the extraction of knowledge from a huge amount of data. [13, 14]

Proposed System

Cancer which is one of the deadliest diseases in today's world has an effective way of reduction in its earliest stages. Its cure rate depends upon its time of detection. Many works have been going on worldwide, but each work lacks in many aspects such as intelligent prediction and inefficiency in implementing the Machine Learning based cancer prediction system. The main intent of the paper is to propose a cancer prediction system that can predict the earliest stage by analyzing the minute set of attributes selected from the dataset.

In this paper, the constructed expert system named the cancer prediction system predicts cancer types (liver, thyroid, leukaemia, lymphoma, lung) which helps to predict cancer type also saving cost and time. Here considered the feature set of symptoms that includes lump area, pain region, swelling area, weight loss, appetite change, fever etc., and predict the class label to which the symptoms of an individual belongs to Lung, Liver, Leukaemia, Lymphoma, Thyroid, No cancer as the class labels. In our dataset, we will be filling the missing values by using mean (shown in figure 4), Calculating the non-missing value means in a column and replacing the missing values of each column separately independent from the others shown in figure 5 which can only be used with numeric data. Accuracy can be predicted by the percentage of correctly classified instances.

$$\text{Accuracy} = (tp + tn) / (tp + tn + fp + fn)$$

where tp, fn, fp and tn represent the number of true positives, false negatives, false positives and true negatives respectively.

Recall is calculated as the ratio of the number of true positives divided by the sum of the true positives and the false negatives.

$$\text{Recall} = \text{True Positive} / (\text{True Positive} + \text{False Negative})$$

$$= \text{True Positive} / \text{Total Actual Positive}$$

The Roc curve or Receiver Operating Characteristic curve is a graphical representation that explains the diagnostic ability of a binary classifier system. Once the user enters the cancer prediction system, they need to provide symptoms. Then the prediction system analyzes the symptoms and displays the cancer type as shown in figure 2.

The cancer prediction system predicts the cancer type of the person based on the symptoms entered by the user. The proposed system uses a logistic regression classifier for training a machine learning model, which takes the symptoms from the user. Here we are adopting a logistic regression algorithm it works on the Data set (shown in figure 3) for training the machine learning supervised model which is used to predict the class label. Based on the class label predicted cancer type appear. Firstly consider a cancer dataset and select a classifier that has high accuracy level and recall score. Then we use that classifier for training and testing. The entered symptoms are recorded and according to them predict the cancer type. This Proposed system helps in the detection of a person's tendency of cancer before going for clinical and lab tests which is costly and time-consuming. This proposed System generates accurate results which can be regarded with a good performance.

Adopted Logistic Regression statistical model is popular which is used for binary classification (example Yes or No, 0 or 1, etc.,) that is for predictions of the types. This is also used for multiclass classification. The hypothesis of logistic regression tends to limit the cost function between 0 and 1. The recall function is used to calculate the ratio of the number of true positives divided by the sum of the true positives and the false negatives. A true positive is an outcome where the model correctly predicts the positive class. A false negative is an outcome where the model incorrectly predicts the

negative class. A ROC (Receiver Operating Characteristic) curve is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters True Positive Rate and False Positive Rate shown in figure 6.

Some of the main modules which are involved are

Accuracy-score (y_test,y_pred)

Recall-score (y_test, y_pred)

Roc-auc-score(y_test, y_pred)

Logistic Regression()

predict()

CONCLUSION

In this world, Cancer becomes a catastrophe for a human being who is suffering from it. Now a day's cancer is a tedious infection in the world. The most successful way to decrease cancer death is to identify it in the early stage. The premature identity of cancer can help cure the illness. So the latest technologies are used to detect the happening of cancer in the premature stage is growing. The main aim of this paper is to identify cancer type based on symptoms given by the user. Here we adopted a supervised learning algorithm and then used the Logistic Regression based on accuracy and recall score i.e., the algorithm which obtains high accuracy level and recalls score. In future, we are going to extend this work by finding the cancer stage and recommending different hospitals and doctors for the particular type of cancer. The advantages of the proposed system are executed with good performance because it generates accurate results.

ACKNOWLEDGEMENT

Authors acknowledge the immense help received from the scholars whose articles are cited and included in references of this manuscript. The authors are also grateful to authors/editors/publishers of all those articles, journals and books from where the literature for this article has been reviewed and discussed

Conflict of Interest: The authors declare that they have no conflict of interest.

Source of Funding: Not Applicable

Authors' Contribution: The author confirms sole responsibility for the following: study conception and design, data collection, analysis and interpretation of results, and manuscript preparation.

REFERENCES

1. Roseline Jecintha I, Poonguzhali. Study on Data Mining Techniques for Cancer Prediction System. Int J Data Mining Techn Appl. 2018; 07(1): 60-63.
2. Malarvizhi. K, Rajivsuresh kumar G. An Instant Guidance on Cancer Prediction and Care Using Web Application. Int J Innov Techn Expl Engg. 2019; 8 (6S): 225-228.
3. Gousbi B, Mohamed Shanavas A R. A Study: Breast Cancer Prediction Using Data Mining Techniques. Asi J Comp Sci Tech. 2019; 8 (S2), 52-56.
4. Priyanga A, prakasam S. The Role of Data Mining-Based Cancer prediction system (DMBCPS) in Cancer Awareness. Int J Compt Sci Engg Commun. 2013; 1(1): 381.
5. Samiksha Zaveri, Kamini Solanki. Data Mining Technique Used For Diagnosis and Prognosis of Cancer Disease. J Emerg Techn Innov Res. 2018; 5(11)
6. Eshlaghy, A.T, Poorebrahimi A, Ebrahimi M, Razavi A. R, Ahmad L G. Using three machine learning techniques for predicting breast cancer recurrence. J Heal Med Inform. 2013; 4(2): 124
7. Konstantina Kourou, Themis P.Exarchos, Konstantinos P.Exarchos, Michalis V.Karamouzis, Dimitrios I.Fotiadis. Machine learning applications in cancer prognosis and prediction. Omputat Str Biotech J. 2015; 13: 8-17
8. Joseph A. Cruz, David S. Wishart. Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Informatics. 2007; 2: 59-77
9. Nath, A.S pal A, Mukhopadhyay S. A survey on cancer prediction and detection with data analysis. Innov Syst Softw Engg. 2019; 12(5): 185-187 https://doi.org/10.1007/s11334-019-00350-6
10. Yuanjie Zheng, Brad,M., Keller, Shonket Ray, Yan Wang, Emily F. Conant, James C. Gee, Despina Kontos, Parenchymal. Texture analysis in digital mammography: A fully automated pipeline for breast cancer risk assessment. Med Phys. 2015; https://doi.org/10.1118/s1.4921996
11. Chih-Jen Tseng, Chi-Jie LU, Chi-chang chang, Gin-Den chen. Application of Machine Learning to predict the recurrence-Proneness for cervical cancer. Neur Comp Appli. 2014; 21(3): 349-352. https://doi.org/10.1007/s00521-013-1359-1
12. Chi-chang chang, Ssu-Han Chen. Developing a Novel Machine Learning-Based Classification Scheme for Predicting SPCs in Breast Cancer Survivors. Front Gen. 2019; https://doi.org/s10.3389/fgene.2019.00848
13. Maalel, A., Hattab, M. Literature review: Overview of Cancer Treatment and Prediction Approaches based on Machine Learning: Smart Systems for E-Health. Adv Inf Know Proc Springer. 2019; p. 324
14. George D. Magoulas., Andriana Prentza. Machine Learning in Medical Applications 2049; Springer LNCS; 2001.

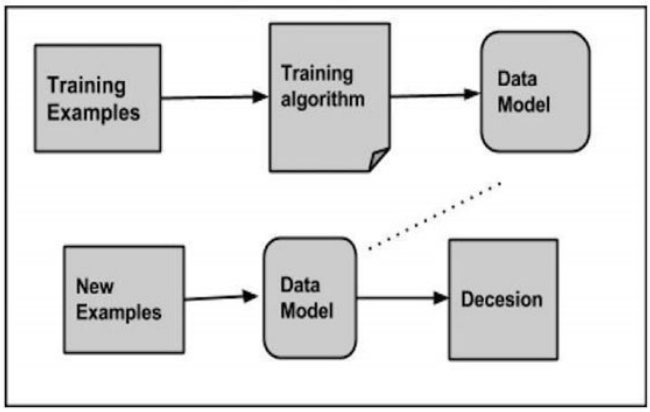

Figure 1: Classification working process.

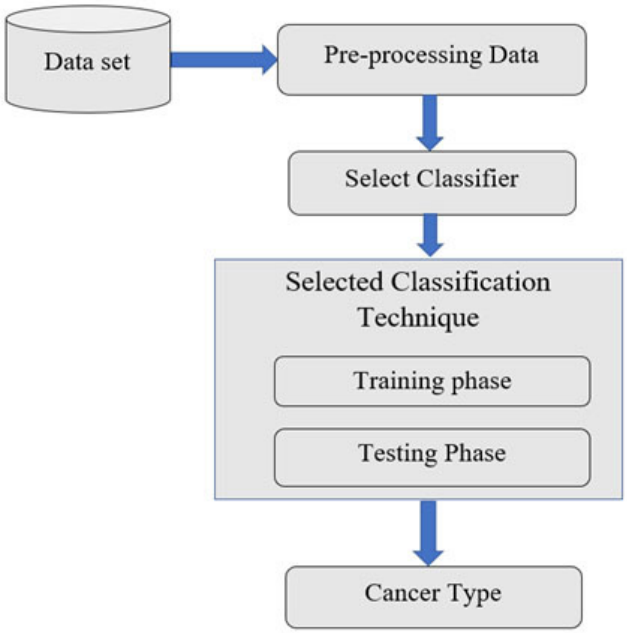

Figure 2: Cancer Prediction system.

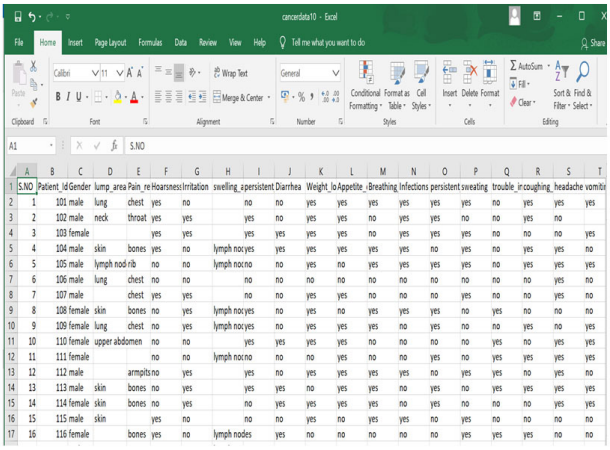

Figure 3: Cancer dataset.

	S.NO	Patient_Id	Gender	lump_area	Pain_region	Hoarsness	Irritation	swelling_area	persistent_fatigue	Diarrhea	...
0	1	101	0	2.000000	1.000000	0	1	0.592593	1.0	1	...
1	2	102	0	0.000000	0.000000	0	0	0.592593	0.0	1	...
2	3	103	1	2.013158	3.223684	0	0	0.592593	0.0	0	...
3	4	104	0	1.000000	2.000000	0	1	0.000000	0.0	0	...
4	5	105	0	7.000000	6.000000	1	1	0.000000	1.0	1	...
...	...	...	...	...	...	...	...	...	...	...	...
95	96	196	0	1.000000	2.000000	0	1	0.000000	0.0	0	...
96	97	197	0	3.000000	8.000000	1	1	0.592593	0.0	0	...
97	98	198	0	4.000000	3.223684	1	1	0.592593	0.0	0	...
98	99	199	1	2.000000	1.000000	0	0	0.000000	1.0	0	...
99	100	200	0	2.000000	1.000000	0	1	0.592593	0.0	1	...

100 rows × 29 columns

Figure 4: Filling Missing Values.

	S.NO	Patient_Id	Gender	lump_area	Pain_region	Hoarsness	Irritation	swelling_area	persistent_fatigue	Diarrhea	...
0	1	101	0	2.0	1.0	0	1	NaN	1.0	1	...
1	2	102	0	0.0	0.0	0	0	NaN	0.0	1	...
2	3	103	1	NaN	NaN	0	0	NaN	0.0	0	...
3	4	104	0	1.0	2.0	0	1	0.0	0.0	0	...
4	5	105	0	7.0	6.0	1	1	0.0	1.0	1	...
...	...	...	...	...	...	...	...	...	...	...	...
95	96	196	0	1.0	2.0	0	1	0.0	0.0	0	...
96	97	197	0	3.0	8.0	1	1	NaN	0.0	0	...
97	98	198	0	4.0	NaN	1	1	NaN	0.0	0	...
98	99	199	1	2.0	1.0	0	0	0.0	1.0	0	...
99	100	200	0	2.0	1.0	0	1	NaN	0.0	1	...

Figure 5: Replaced Data.

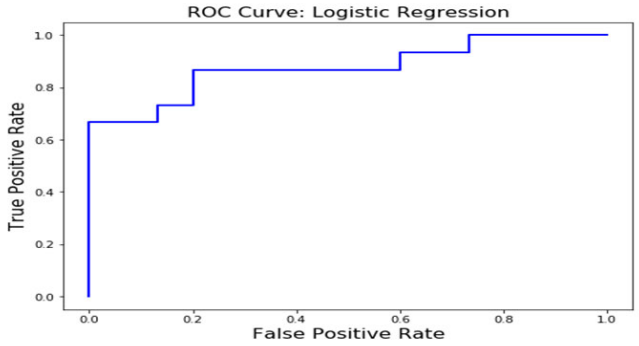

Figure 6: ROC curve.