PSYCHOLOGY AND EDUCATION (2021) 58(1): 6790-6795
ISSN: 1553-6939

BREAST CANCER DETECTION WITH MACHINE LEARNING

Erugu Krishna[1],B.Venkateswarlu[2],Dr.M.Bal Raju[3],Dr.Ch Ramesh Kumar[4],
Assistant Professor[1,2],Professor[3,4],
Department of CSE,
Pallavi Engineering College[1,2,4],Swamy Vivekananda institute of Technology[3]
Mail ID:krishna.cseit@gmail.com,Mail ID:bvenkat1109@gmail.com,Mail ID:drrajucse@gmail.com
Kuntloor(V),Hayathnagar(M),Hyderabad,R.R.Dist.-501505.

Abstract

Accordin to the Breast Cancer Institute (BCI), breast cancer is one of the most dangerous forms of cancers that, if diagnosed and treated early enough, may be successfully treated for women all over the world. It is believed by medical specialists that detecting this cancer in its early stages can help save people's lives by preventing it from spreading. This website, which covers more than 120 distinct types of cancer and the genetic disorders that are connected with them, provides personalised therapy suggestions based on the individual's medical history. Machine learning algorithms are used to detect the vast majority of breast cancers, which accounts for the majority of cases. This paper presents an adaptive ensemble voting approach for newly diagnosed breast cancer that was developed using the Wisconsin Breast Cancer database and is based on a randomised controlled experiment that was conducted using the Wisconsin Breast Cancer database. The Wisconsin Breast Cancer database was used in the research for this paper. The goal of this research is to compare and explain how the ANN and logistic algorithms, when used in conjunction with ensemble machine learning algorithms for diagnosing breast cancer, generate greater outcomes when the number of variables is reduced. The Wisconsin Diagnosis Breast Cancer dataset, which was produced specifically for this study, was used in this investigation. For the sake of comparison, this study is being compared to other comparable studies that have previously been published. When ANN methodology and the logistic algorithm are coupled, they provide a classification accuracy rate of 98.50 percent when compared to another machine learning technique, as demonstrated by a comparison to another machine learning strategy (Figure 1).

1. INTRODUCTION

World-wide, cancer ranks first among all diseases in terms of mortality, with breast cancer ranking first among all cancers in terms of hazard to women. According to the American Cancer Society, breast cancer claims the lives of hundreds of people every year. The physical detection of breast cancer is time-consuming, and it is difficult for the physician to determine what stage of the sickness is being dealt with at any one moment. As a result, in recent years, the detection of cancer through the use of various automated diagnostic technologies has gained growing importance and importance. In order to detect breast cancer, a range of algorithms and techniques are available, including the Support Vector Machine, Nave Bayes, Kernel Neural Network, and Convolution Neural Network, among others. It is the most recent algorithm in deep learning, and it is also the most recent algorithm in deep learning that is also utilised for classification. It is also the most recent algorithm in deep learning. In

deep learning, it is the most current algorithm to be developed, and it is also the most recent algorithm to be developed in deep learning. The categorization and identification of objects in photographs is mostly performed through the use of CNN and deep learning algorithms, which are becoming increasingly popular. In this study, we make use of the open database maintained by the University of California, Irvine (UCI), which contains two classes of tumours: one that is benign and another that is malignant, where benign Tumor is a non-cancerous tumour and malignant Tumor is a cancerous tumour, for training and testing purposes. There are still a large number of researchers working on this topic today, with the goal of discovering and diagnosing cancer in its earliest stages as soon as feasible. Early-stage cancer is less severe and expensive to treat than later-stage cancer, which is why many researchers are now concentrating their efforts on developing a reliable diagnostic system that can detect tumours at the earliest possible stage in their progression and treat them as effectively as possible. Therefore, therapy can begin sooner and the rate of remission can be increased as a result, both of which are beneficial to the patient. The primary goal of this project is to conduct a comparative assessment of a variety of different machine learning approaches, which will be accomplished through the use of Artificial Neural Networks. The remainder of the paper may be found in the parts that follow this one: Among the sections of the proposed research that contain a review of relevant literature is Section 2, which is included in the proposal itself. This section describes the architectural design of the planned building project in greater depth than the preceding sections. Section 4 discusses the strategy that will be utilised for the proposed study, which will be explained in greater detail later in this paper. Section 5 discusses the methodology that will be used for the proposed study. Section 5 offers a synopsis of the findings of the proposed investigation. Section 5 delves into the process of determining the qualities that will be used for the planned task in greater depth. Section 6 provides a summary of the method. On the sixth page, we will go through the procedures that will be followed in order to complete the assignment that has been assigned to us. A description of the results is provided in Section 7, and the scope of the planned study is concluded in Section 8. Section 7:

Literature survey

At DeSE 2016, the 9th International Conference on Developments in eSystems Engineering (DeSE), which took place in Liverpool in 2016, M. R. Al-Hadidi, A. Alarabeyyat, and M. Alhanahnah presented their findings. It was published at DeSE 2016, the 9th International Conference on Developments in eSystems Engineering (DeSE), which was held in October. The study, "Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm," was written by a team of researchers. As normal clinical practise in the medical field when it comes to diagnosing breast cancer, breast cancer detection pictures are used to aid in the identification of the disease. Due to the widespread use of digital mammography, breast cancer and other abnormalities in women of reproductive age are being diagnosed sooner and more successfully than ever before. Digital mammography is now the most widely used screening tool in the world. As texture descriptors, we propose the use of Polar complex Exponential Transform (PCET) moments to automatically differentiate between malignant and non-malignant breast tissues in a Computer-Aided Detection (CAD) system. Several pre-processing procedures are performed prior to displaying the image on the screen, including histogram ROI selection, which is used to define the input Region of Interest (ROI) and histogram ROI selection. It is then possible to extract features from the data by using the PCET moments that have been calculated. It is being utilised to increase the accuracy of the categorization system in the CAD system. In order to increase the classification accuracy of the CAD system, a sophisticated classifier known as the Adaptive Differential Evolution Wavelet Neural Network (ADEWNN) is utilised.

There was a conference in Kuala Lumpur, Malaysia, called the IEEE 10th International Colloquium on Signal Processing and its Applications. The proceedings of the conference are now accessible online. "Breast cancer mass localization based on machine learning" was the title of a paper presented at the IEEE 10th International Colloquium on Signal Processing and its Applications in Kuala Lumpur, Malaysia, in 2014. The study was written by A. Qasem and colleagues and was presented in 2014.

BIRADS is an abbreviation for Breast Imaging, Reporting, and Data System, which is used in the field of breast imaging. To increase patient safety, a system has been developed to standardise mammography reports and reduce ambiguity during the interpretation of mammogram images. The categorising of BIRADS is one of the most difficult jobs a radiologist needs to accomplish throughout the course of his or her professional life. In some cases, an oncologist may be able to prescribe the most suitable treatment for a specific patient after gathering adequate information during the BIRADS staging process. This research sought to construct a model that could be used to categorise BIRADS based on mammography pictures and reports, which was one of the main goals of the project. To test the model, an autonomously generated set of rules will be applied. Type-2 fuzzy logic as a classifier will be implemented within the model, and will be used to classify data within the model. Calculations will be performed on the suggested model's accuracy, specificity and sensitivity, and the results will be compared against rules established by experts in the field to assess whether the model is successful. The research is divided into many sections, the first of which is the collecting of data from the Radiology Department of the Hospital of the National University of Malaysia. The second section is the analysis of the data collected. The second phase involves the examination of the information gathered (UKM). Data was first cleaned up to remove noise and gaps, and then it was subjected to statistical analysis to determine its significance. Next that, an algorithm was created by selecting type-2 fuzzy logic and applying it to the Mamdani model, which was completed in the following phase. During their inquiry, members of the study team used three different sorts of membership functions, each of which was unique. Furthermore, in addition to rules derived from experts, the model also depended on rules generated automatically by the system through the use of rough set theory. All of these principles were taken into consideration when developing the model's computations. Finally, the model was put through its paces and given specific instructions on how to get the best possible results. Precision rates of 78 percent are obtained by using expert rules, but accuracy rates of 89 percent are obtained by using triangular membership functions based on rough specified rules According to the use of expert rules, the sensitivity is 98.24 percent, but the sensitivity achieved with the application of rough established rules is 93.94 percent.

3. SYSTEM ANALYSIS

Existing system

Machine learning technologies are available for the prediction and diagnosis of breast cancer, and they may be utilised in concert with one another to provide the most accurate results possible. In machine learning, there are many distinct types of algorithms, including the Artificial Neural Network (ANN), Naive Bayes, Support Vector Machine (SVM), KNearest Neighbor (KNN), and Convolutional Neural Network (CNN) (CNN). Many researchers have used various data sets, including the SEER data set, the Wisconsin data set, and multiple hospital-based datasets, among other sources, to conduct breast cancer research studies. Mammogram pictures, as well as the SEER data set and the Wisconsin data set, have all been used as datasets in this study. A variety of items may be extracted and selected from these databases, allowing authors to finish their study assignment. The following are some important research reviews to take into consideration. Using the DWT and the BPNN tools for picture filtering and processing, author Moh'dRasoul was able to attain an overall accuracy of 93.7 percent. Using the WHAVE algorithm in conjunction with the Wisconsin Breast Cancer Database, author Clemen Deng was able to predict the presence of breast cancer with 99 percent accuracy. Using the marker Controller Watershed technique, it took AshwaqQasem three months to attain 95 percent accuracy, and the results were remarkable. In his research, the author, AlirezaIsareh, employed signal to noise ratio in conjunction with SVM to achieve an accuracy of 98.80 percent, according to the results. Author Junaid Ahmad employed Adaptive Resonance Theory in conjunction with the University of California, Irvine database to achieve an accuracy of 84.21 percent, according to the results of his research. BM Gayathri's work on a comparison assessment of relevance vector machines, according to her, yielded a classification accuracy rate of 97 percent based on a comparative evaluation of relevance vector machines. A set of 150 images was analysed by the author, Ms.H.R. Mhaske, who utilised KNN and SVN to identify items in the photographs. Using these techniques, she was able to obtain an accuracy of 80-90 percent. According to his findings, the author S. Arunautilised Nave Bayes and SVM with the UCI database to get accuracy ranging from 68 to 79 percent using the UCI database. YohannesTsehay designed a poorly supervised computer-assisted detection system that was used to collect biopsy data in order to achieve the aim of learning from the data. SudarshanNayak conducted his research with the help of Nave Bayes and SVM, and he was able to reach a 98% accuracy rate. According to his study, Ryota Shimizu was able to reach 90 percent accuracy through the application of Deep Learning and neural networks. Other researchers, as well as a large number of other writers, have conducted research on the detection and diagnosis of breast cancer using a variety of machine learning techniques.

Disadvantages:

• A lower overall grade due to inaccuracy
performance that falls short of the required standard
• It is impossible to foresee how the litigation will turn out.

Proposed system

We'll go over the recommended approach for computing results in this part, so pay close attention! I started by downloading the Wisconsin Breast Cancer Database from the University of California, Irvine website and pre-processing the data to extract 16 critical features about breast cancer. For the most important characteristics, we employed the Recursive feature Elimination Algorithm with the Chi2 approach, which is a recursive feature elimination algorithm. The Chi2 technique enabled us to come up with 16 of the most useful features for our product. As a result, each of the ANN and Logistic algorithms was tested independently, and the accuracy of each method was determined. In the end, we used the Ensemble Voting procedure that was advised to us in order to establish the most effective method of detecting breast cancer disease. Using a neural network and a logistic algorithm, we have developed an ensemble technique for the detection of breast cancer in this study. This approach has been shown to be successful. In all, three phases are included in the procedure: data pre-processing, feature selection, and the development of a voting model for each category. As a basis for our inquiry, we employed the BCI dataset, which consists of 569 rows and 30 columns of information. Specifically, the BCI dataset was utilised in this work. To begin, we examined and then evaluated the properties of the default dataset, as we mentioned in more detail in the experiment section. The Univariate Features Selection approach and the Recursive Features Selection method with Cross Validation Method were both utilised to choose features for the analysis. Both techniques were effective, and we found them to be useful.

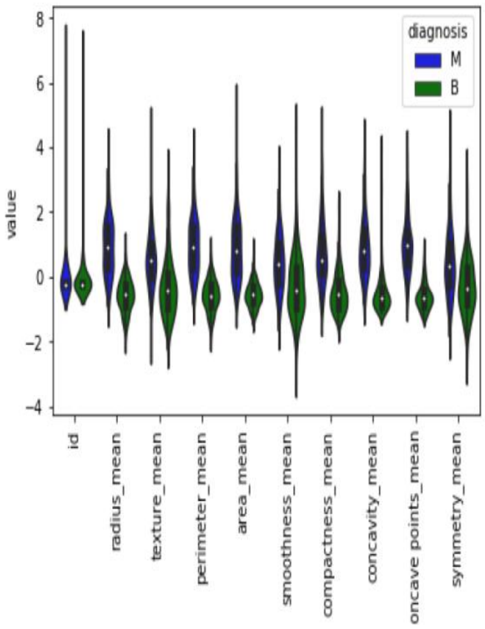
The following information is provided: A. Pre-processing information Data pre-processing is a technique used in data mining that consists of turning raw data into a format that can be readily understood by the user. In many cases, real-world data is insufficient, inconsistent, and lacking in particular behaviours, and as a result, it is likely to contain a lot of inaccuracies as a result of these limitations. In prior research, it was demonstrated that pre-processing data was a viable approach of overcoming such challenges. Performing data pre-processing on raw data in order to prepare it for further processing is referred to as data preparation (or data preparation). The standardisation approach was used to pre-process the UCI dataset for the aim of pre-processing it for further analysis. B. The Use of Standardization as a Strategist: It is referred to as the standardisation approach since the dataset serves as a common criterion for a range of machine learning estimators in this method, which is why it is known as such. Several data visualisations are presented in this paper, which may be utilised for data pre-processing. We must first count the number of malignant and benign tumours in the full dataset and put them on a graph so that they can be easily seen. Second, a Violin plot was created, which was then utilised to aid in the data reduction process. In order to examine the distributions of distinct variables across multiple levels of a category variable, there are many alternative methods to express the distribution of quantitative data across different levels of a category variable. A scatterplot with dots that do not overlap one another is then generated as a result of this procedure. Comparing this method to the one used previously, it provides a more accurate representation of the distribution of data. The link between dataset attributes has been illustrated using a relationship scatterplot in order to help in better knowledge of the association. The scatterplot of links between the characteristics of the dataset was represented by this graph.

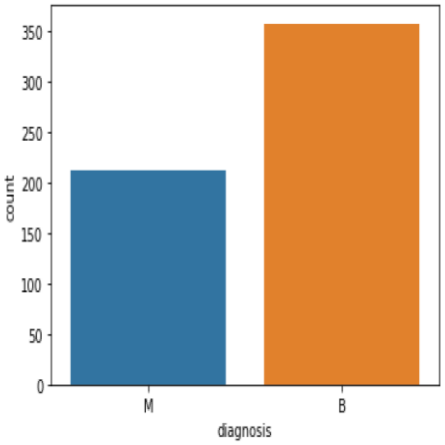
Advantages:

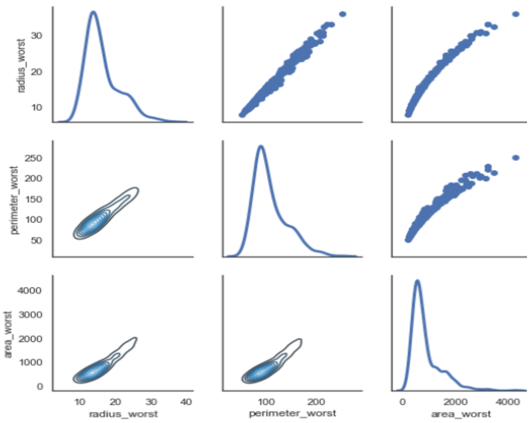
- Exceptional accuracy and overall effectiveness

Make an informed guess about the bigger resolution and use it to determine the smaller resolution.

4. Results:







5. Conclusion

In this study, we offer an ensemble machine learning approach for identifying breast cancer that has been validated. This approach has an accuracy rate of 98.50 percent, which is exceptional, as seen in the following table and graph. To make a cancer diagnosis in this study, we used just 16 variables that were gathered from the participants. Throughout the future, we intend to experiment with all of the components of UCI in order to achieve the maximum degree of accuracy possible. Neuronal networks are also excellent for analysing human vital data, as proved by our research, and we are able to do prediagnosis without the requirement for specialised medical skills..

REFERENCES

[1] M. R. Al-Hadidi, A. Alarabeyyat and M. Alhanahnah, "Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm," 2016 9th International Conference on Developments in eSystems Engineering (DeSE), Liverpool, 2016, pp. 35-39.

[2] C. Deng and M. Perkowski, "A Novel Weighted Hierarchical Adaptive Voting Ensemble Machine Learning Method for Breast Cancer Detection," 2015 IEEE International Symposium on Multiple-Valued Logic, Waterloo, ON, 2015, pp. 115-120.

[3] A. Qasem et al., "Breast cancer mass localization based on machine learning," 2014 IEEE 10th International Colloquium on Signal Processing and its Applications, Kuala Lumpur, 2014, pp. 31-36.

[4] A. Osareh and B. Shadgar, "Machine learning techniques to diagnose breast cancer," 2010 5th International Symposium on Health Informatics and Bioinformatics, Antalya, 2010, pp. 114-120.

[5] J. A. Bhat, V. George and B. Malik, "Cloud Computing with Machine Learning Could Help Us in the Early Diagnosis of Breast Cancer," 2015 Second International Conference on Advances in Computing and Communication Engineering, Dehradun, 2015, pp. 644- 648.

[6] B. M. Gayathri and C. P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2016, pp. 1-5.

[7] H. R. Mhaske and D. A. Phalke, "Melanoma skin cancer detection and classification based on supervised and unsupervised learning," 2013 International conference on Circuits, Controls and Communications (CCUBE), Bengaluru, 2013, pp. 1-5.

[8] S. Aruna, S. P. Rajagopalan and L. V. Nandakishore, "An algorithm proposed for Semi- Supervised learning in cancer detection," International Conference on Sustainable Energy and Intelligent Systems (SEISCON 2011), Chennai, 2011, pp. 860-864.

[9] Y. Tsehay et al., "Biopsy-guided learning with deep convolutional neural networks for Prostate Cancer detection on multiparametric MRI," 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), Melbourne, VIC, 2017, pp. 642-645.

[10] S. Nayak and D. Gope, "Comparison of supervised learning algorithms for RF-based breast cancer detection," 2017 Computing and Electromagnetics International Workshop (CEM), Barcelona, 2017, pp. 13-14.doi: 10.1109/CEM.2017.7991863