Journal of Engineering Sciences
Vol 11, Issue 6,June/2020
ISSN NO:0377-9254

Identification of Breast Cancer Stages using Machine Learning Techniques

Dhanushiya R.[1], Divyabharathi S.[2], Harinni S.[3], Dr.L.Jaba Sheela[4]

[1]UG Student, BE Computer Science and Engineering, Panimalar Engineering College, Chennai
[2]UG Student, BE Computer Science and Engineering, Panimalar Engineering College, Chennai
[3]UG Student, BE Computer Science and Engineering, Panimalar Engineering College, Chennai
[4]Supervisor, Dept of Computer Science and Engineering, Panimalar Engineering College, Chennai

ABSTRACT:

Breast cancer is one of the most common cancer in women worldwide, an invasive tumor that develops in the mammary gland. Our analysis provides a comprehensive guide to sensitivity analysis of model parameters with regard to performance in detection of breast cancer stages by predicting result in the form of dataset attributes. Using machine learning methods for diagnostic can nominally increase the processing speed and on a big scale can make the diagnostic cheaper. The analysis of dataset to capture several information's like, variable identification, uni-variate analysis, bi-variate analysis, missing value treatments etc is done using supervised machine learning algorithm. The main objective is to predictive analytics model to diagnose breast cancer stages of patients. Additionally, we discuss the performance from the given dataset with evaluation classification report and identify the confusion matrix. The data validation, preparing and visualization will be applied on the entire given dataset. So, aim of categorizing data from priority and the result shows that the effectiveness of the proposed machine learning algorithm technique can be compared with best accuracy, precision, Recall and F1 Score.

Keywords: Dataset, Python, Prediction of Accuracy result.

INTRODUCTION

Machine learning (ML) is to predict the long run from past data and it's a sort of AI (AI) that gives computers with the power to be told without being explicitly programmed. Machine learning focuses on the event of Computer Programs which will change when exposed to new data and therefore the basics of Machine Learning, implementation of machine learning algorithm using python.

Techniques of Supervised Machine Learning algorithms include logistic regression, Decision Trees and support vector machines etc. A classification model is made to predict breast carcinoma from its feature. The dataset used for analysis contains many inconsistencies like missing values, outliers and it should be handled before being employed to create the model. Breast carcinoma is cancer that forms within the cells of the breasts. To provide a support for breast cancer awareness and to find the appropriate diagnosis and treatment of breast cancer. Cancer survival rates have increased, and therefore the number of deaths related to this disease is steadily declining, largely because of factors like earlier detection, a replacement personalized approach to treatment and a far better understanding of the disease.

Signs and symptoms of breast cancer may include:

- A breast lump or thickening that feels different from the encircling tissue .
- Change within the dimensions, shape or appearance of a breast .
- Changes to the heal the breast, like inflammation.

- A newly inverted nipple.
- Peeling, scaling, crusting or flaking of the coloured area of skin around the nipple or breast skin.
- Redness or pitting of the heal your breast, rather just like the skin of an orange

LITERATURE SURVEY

1. Title: Comparison of Machine Learning Methods for Breast Cancer Diagnosis

Author: Ebru Aydındag Bayrak, Pınar Kırcı,Tolga Ensari

Year: 2019

Breast Cancer is the most frequent disease as a cancer type for women. Therefore, any development for diagnosis and prediction of cancer disease is capital important for a healthy life.

It discussed two popular machine learning techniques for Wisconsin Breast Cancer classification. Artificial Neural Network and Support Vector Machine are used as ML techniques for the classification of WBC (Original) dataset in WEKA tool.

The effectiveness of applied ML techniques is compared in term of key performance metrics such as accuracy, precision, recall and ROC area. Based on the performance metrics of the applied ML techniques, SVM (Sequential Minimal Optimization Algorithm) has showed the best performance in the accuracy of 96.9957 % for the diagnosis and prediction from WBC dataset.

2. Title:Breast Cancer Detection Using Extreme Learning Machine Based on Feature Fusion with CNN Deep Features

Author: Zhiqiong Wang, Mo Li, Huaxia Wang

Year : 2019

The accuracy of existing CAD systems remains unsatisfactory. This paper explores a breast CAD method supported feature fusion with Convolution Neural Network (CNN) deep features. First, we propose a mass detection method supported CNN deep features and Unsupervised Extreme Learning Machine (US-ELM) clustering. Second, we build a feature set fusing deep features, morphological features, texture features, and density features. Third, an ELM classifier is developed using the fused feature set to classify benign and malignant breast masses. Early detection of lumps can effectively reduce the death rate of carcinoma. The mammogram is widely utilized in early screening of carcinoma because of its relatively low expense and high sensitivity to minor lesions. within the actual diagnosis process, however, the accuracy is negatively laid low with many factors, like radiologist fatigue and distraction, the complexity of the breast structure, and therefore the subtle characteristics of the early-stage disease. The computer-aided diagnosis (CAD) for carcinoma can help address this issue. Although the old diagnosis method has been widely used, its accuracy still has to be improved. The standard of the handcrafted feature set directly affects the diagnostic accuracy, and hence an experienced doctor plays a awfully important role within the process of manual feature extraction. Its main idea is to use deep features extracted from CNN to the 2 stages of mass detection and mass diagnosis. Within the stage of mass detection, a method based on sub-domain CNN deep features and US-ELM clustering is developed. In the stage of mass diagnosis, an ELM classifier is utilized to classify the benign and malignant breast masses using afused feature set, fusing deep features, morphological features, texture features, and density features. In the process of breast CAD, the choice of features is the key in determining the accuracy of diagnosis.

3.Title : Breast Cancer Diagnosis in Digital Breast Tom synthesis: Effects of Training

Sample Size on Multi-Stage Transfer Learning using Deep Neural Nets

Author: Ravi K. Samala, Heang-Ping Chan and Lubomir Hadjiiski

Year : 2018

In a CNN, the convolutional layers near the input are generic and the deeper layers are specific to the target task. Transfer learning from one domain (e.g., non-medical images) to another (e.g., medical images) is to utilize these generic features while transforming or fine-tuning the deeper features to a target task. However, when the available training data from the target domain are limited, the pre-trained features may not be sufficiently fine-tuned to the

target task. Instead of transfer learning directly to the target domain with small training set, additional, intermediate stages of transfer learning from related auxiliary domains may help improve learning in the target task. Using cross validation, we selected the best transfer network from six transfer networks by varying the level up to which the convolutional layers were frozen. In a single-stage transfer learning approach, knowledge from CNN trained on Image Net data was fine-tuned directly with DBT data. In a multi-stage transfer learning approach, knowledge learned from Image Net was first fine-tuned with the mammography data and then fine-tuned with the DBT data. Two transfer networks were compared for the second stage transfer learning by freezing most of the CNN structure versus freezing only the first convolutional layer.

Finally, it shows that the limited data availability in a target domain can be alleviated with pre-training of the CNN using data from similar auxiliary domains. We also show that the gain in CNN performance from the additional stage of fine-tuning with the auxiliary data depends on the relative sizes of the available training samples in the target and the auxiliary domains, and proper selection of the transfer learning strategy. Reporting the best performance through exhaustive search using a ”test” set can be overly optimistic. It is therefore important to validate the generalizability of the trained CNN with independent unknown cases. TO utilize deep convolution neural networks (CNNs) for pattern recognition tasks in medical imaging, transfer learning is commonly used due to the lack of large training data. Our work demonstrates that multi-stage transfer learning can take advantage of the knowledge gained through source tasks from unrelated and related domains. We show that the limited data availability in a target domain can be alleviated with pre-training of the CNN using data from similar auxiliary domains. We also show that the gain in CNN performance from the additional stage of fine-tuning with the auxiliary data depends on the relative sizes of the available training samples in the target and the auxiliary domains, and proper selection of the transfer learning strategy.

4. Title : A multimodal deep neural network for human breast cancer prognosis prediction by Integrating multi-dimensional data

Author: Dongdong Sun, Minghui Wang

Year : 2018

BREAST cancer is the most highly aggressive cancer and a major health problem in females, and a leading cause of cancer-related deaths worldwide. The ability of predicting cancer prognosis more accurately not only could help breast cancer patients know about their life expectancy, but also help clinicians make informed decisions and further guide appropriate therapy. To comprehensively evaluate our proposed method, we use ten-fold cross validation experiment in consistent with previous existing studies of cancer prognosis prediction. Specifically, the patients in our experiment are randomized into ten subsets. For every round, nine of those ten subsets are further divided into training (80%) and validation (20%) sets, while the remaining one subset is utilized as testing set. In this way, we obtain the prediction scores of each testing subset after ten rounds and then merge them as an overall prediction scores. Besides, in our study, MDNNMD does not optimize the model configurations and weight coefficients simultaneously. For performance evaluation, we plot receiver operating characteristic (ROC) curve, which shows the interplay between sensitivity and 1-specificity by varying a decision threshold, and computes the AUC. The evaluation metric, Sensitivity (Sn), Specificity (Sp), Accuracy (Acc), Precision (Pre) and Matthew’s correlation coefficient (Mcc) are also used for performance evaluation. We compare the performance of MDNNMD with three widely used methods for prognosis prediction of breast cancer: SVM, RF and LR. Ten-fold cross validation experiment for prognosis prediction of breast cancer is conducted with four different methods. In this study, we use an RF and LR package obtained from scikitlearn. In this work, we present a novel multimodal deep neural network by integrating multi-dimensional data named MDNNMD to predict the survival time of human breast cancer. To efficiently incorporate multidimensional data including gene expression profile, CNA and clinical data in breast cancer, three independent DNN models are constructed to generate

a final multimodal DNN model considering the heterogeneity of different types of data.

5. Title : Clinical Pilot Application of Super-resolution US Imaging in Breast Cancer

Author: S. Dencks, M. Piepenbrock and T. Opacic

Year : 2018

To compare the results of state-of-the-art contrast-enhanced to super-resolution US imaging and systematically analyze the measurements to get indications for the improvement of image acquisition and processing in future clinical studies. In this regard, the application of a saturation model to the reconstructed vessels is shown to be a valuable tool not only to estimate the measurement times necessary to adequately reconstruct the microvasculature but also for the validation of the measurements. The clinical interest in the microvasculature of tissues is manifold because deviations from normal vessel growth play a role in numerous diseases, like inflammatory or cancers or, blinding eye diseases. Particularly for tumors, it is known that their vascularization is morphologically abnormal and that features like the tortuosity of vessels, their branching, their irregular vessel diameters, and the inhomogeneity throughout the tumor contain vital information on its aggressiveness. Functional parameters like the time-to-peak, peak enhancement and upslope of conventional time-intensity (TI) curves are limited to a global interpretation of perfusion. By revealing vascular features at super-resolution and quantifying even very low flow velocities of single vessels, ULM is expected to substantially improve the differential diagnosis, prognostication, and the monitoring and prediction of therapy responses. However, the potential of ULM is strongly interrelated with the technical feasibility in a clinical set-up.

It could show that clinical super-resolution imaging is feasible with a single contrast agent injection within measurement times of less than 5 minutes. Although vessel trees were not imaged completely with the statistical sampling by the MB, relevant parameters could be derived also from incomplete vessel trees by investigating their reconstruction over time.

6.Title : Identification of Novel Scaffolds with Dual Role as Antiepileptic and Anti-Breast Cancer

Author: Shailima Rampogu1, Ayoung Baek1, Rohit Bavi

Year : 2018

The aim of the present study is to identify novel antiepileptic aromatase inhibitors with higher activity exploiting the ligand-based pharmacophore approach utilizing the experimentally known inhibitors. The resultant Hypo1 consists of four features and was further validated by using three different strategies. Hypo1 was allowed to screen different databases to identify lead molecules and were further subjected to Lipinski‘s Rule of Five and ADMET to establish their drug-like properties. Consequently, the obtained 68-screened molecules were subjected to molecular docking by GOLD. Furthermore, the compounds with the highest dock scores were assessed for molecular interaction. Later, the MD simulation was applied to evaluate the protein backbone stabilities and binding energies adapting GROMACS 5.0.6 and MM/PBSA which was followed by the density functional theory (DFT), to analyze their orbital energies and further the energy gap between them. Eventually, the number of Hit molecules was culled to three projecting Hit1, Hit2, and Hit3 as the potential lead compounds based on their highest dock scores, hydrogen bond interaction, lowest energy gap and the least binding energies and stable MD results.

7.**Title :**Breast tumor detection using empirical mode decomposition features

Author: Hongchao Song, Aidong Men and Zhuqing Jiang

Year: 2017

Principal component analysis (PCA) is one of the most widely used feature extraction methods; however, PCA is negatively impacted by signal misalignment. Empirical mode decomposition (EMD)-based feature extraction method that is more robust to signal misalignment. The experimental results obtained from clinical data indicate that the detection accuracy is improved by the combination of features from EMD and PCA. Explored the use of an EMD-based feature extraction method for microwave breast tumor detection. We were motivated to explore

the use of EMD features because of their potential robustness to the system jitter that is common in microwave breast cancer scans. We evaluated the detection performance using EMD-based features and the commonly used PCA features using clinical trial data collected over an eight-month period combined with a numerical tumor response construction method. We observed that the use of features based on EMD leads to significantly improved detection performance compared to PCA-based features. The motivation for microwave-based breast tumor detection is that significant differences in the dielectric properties exist between malignant and healthy breast tissues in the microwave frequency range. The estimated malignant-to-normal breast tissue contrast is approximately 2:1 to 10:1 depending on the density of the normal tissue. Machine learning algorithms are used to capture the differences that exist between normal and malignant tissues. Feature extraction is an important step in achieving good breast cancer detection performance. Misalignment among breast scans, which is caused by equipment movement and the system's intrinsic jitter in the clock and sampling oscilloscope, results in different features being extracted from different scans of the same patent. Therefore, creating features that are insensitive to the time shift existing among scans is important.

PROPOSED SYSTEM

The main objective of this predictive analytics model is to seek out the stage of breast cancer the patient is affected by in order that it should help the physician to see the diagnosis plan and treatment for that specific stage.

The analysis are often divided into four sections

1. Identifying the information and Data Sources
2. Exploratory Data Analysis
3. Pre-Processing the info
4. Build model to predict cancer stages

1. Identifying the information and Data Sources :

Machine learning needs data gathering have lot of past data. Data gathering have sufficient historical data. Raw data can't be used directly without preprocessing them, it may contain some missing variables, anomalies etc.

Splitting the dataset:

The dataset we use are split into training data and test data. The training set contains a known output and also the model learns on this data so as to be generalized to other data soon. It has the test dataset (or subset) so as to check our model's prediction on this subset and it'll try this using SciKit-Learn library in Python using the train_test_split method. Training and testing this model working and predicting correctly with minimum errors.

2. Exploratory Data Analysis:

Data collection:

The data set collected for predicting patient is split into Training set and Test set. Generally, 7:3 ratios are applied to separate the Training set and Test set. The information model which was created using naive Bayesian algorithm are applied on the Training set and supported the test result accuracy, Test set prediction is completed.

Training the Dataset :

The first imports iris data set which is already predefined in sklearn module and data set is largely a table which contains information about various varieties. For example, to import any algorithm and train_test_split class from sklearn and numpy module to be used during this program. To encapsulate load_data() method in data_dataset variable. Further divide the dataset into training data and test data using train_test_split method. The X prefix in variable denotes the feature values and y prefix denotes target values.

This system divides dataset into training and test data randomly in ratio of 67:33 / 70:30. Then we encapsulate any algorithm. Next, we fit our training data into this algorithm in order that computer can get trained using this data. Now the training part is complete.

Testing the Dataset:

Now, the scale of recent features during a numpy array called 'n' and that we want to predict the species of this features and to try and do using the prediction method which takes this array as input and splits out the predicted target value as output. So, the expected target value comes dead set be 0. Finally to search out the test score which is that the ratio of no.

of predictions found correct and total predictions made and finding accuracy score method which basically compares the particular values of the test set with the expected values.

3. Preprocessing the Information:

Data Wrangling:

This load the info, check for cleanliness, so trim and clean given dataset for analysis. ensure that the document steps carefully and justify for cleaning decisions.

Pre Processing:

The data which was collected might contain missing values which will end in inconsistency. To attain better results data must be preprocessed so on improve the efficiency of the algorithm. The outliers should be removed and also variable conversion must be done. Supported the correlation among attributes it had been observed that attributes that are significant individually include TNM, stages, grade, age, which is the strongest among all. Some variables like applicant income and co- applicant income don't seem to be significant alone, which is strange since by intuition it's considered as important.

4. Building the classification model:

Predicting the cancer problem, decision tree algorithm prediction model is effective due to the next reasons: It provides better ends up in classification problem.

• It is powerful in preprocessing errors, irrelevant variables, and a mixture of contiguous, categorical, and discrete variables.

• It produces out of bag estimate error which has proven to be unbiased in many tests and it's relatively easy to tune with. Tuned model involved by tuned time to time with improving the accuracy.

The significance of stage of the cancer:

The stage of a cancer is also a measurement of the extent of the cancer and its spread. the standard staging system for carcinoma uses a system called TNM.

TNM staging system:

The foremost commonly used tool that doctors use to elucidate the stage is the TNM system. Doctors makes use of th test results and scans to answer these questions:

• **Tumor (T):** How large is the first tumor? Where is it located?

• **Node (N):** Has the tumor distributed across lymph nodes? If so where and also the way many?

• **Metastasis (M):** Has the effects of cancer spread to other parts of the body? If so, where and also the way much?

Tumors are graded between 1 and three.

Grade 1 – the cancer cells look small and uniform like normal cells, and are usually slow-growing compared to other grades of carcinoma

Grade 2 – the cancer cells are slightly larger than normal cells, differ in shape and are growing faster than normal cells

Grade 3 – the cancer cells look different from normal cells, and are usually grows faster than normal cells

Staging:

Staging is employed to assess the dimensions of a tumor, whether it's spread and the way far it's spread. The TNM system is commonly accustomed categorize cancers into four stages.

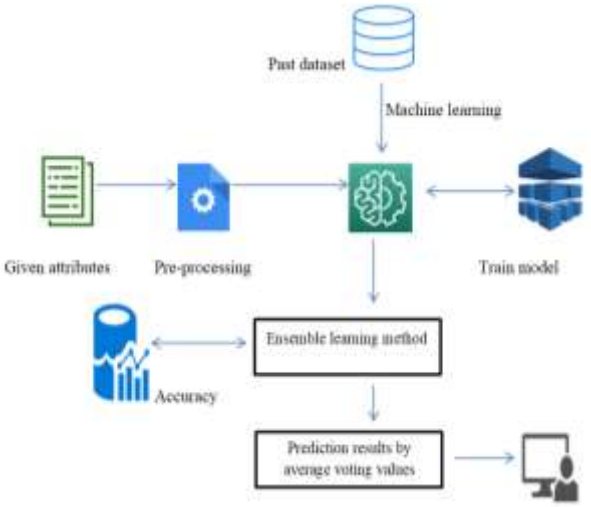
Stage 1: usually implies that a cancer is comparatively small and contained within the breast.

Stage 2: It usually means that the cancer has not yet started to spread around thetissue but the tumor is larger than Stage 1. Sometimes Stage 2 implies that cancer cells have spread into lymph nodes near the tumor.

Stage 3: usually means the cancer is larger. It should have began to spread into surrounding tissues and there are cancer cells within the lymph nodes within the area.

Stage 4: means the cancer has spread from where it began to another body organ. this is often also called secondary or metastatic cancer.

SYSTEM ARCHITECTURE



CONCLUSION

The process of analysis started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. Finding the patient stages and grade with parameter like accuracy, classification report and confusion matrix on public test set of given attributes by supervised machine learning method. So finally, we have built our classification model and to view the machine learning classification algorithm gives the best results for our dataset. Well it's not always applicable to every dataset. Our dataset is always need to analyze for choose our model and then our machine learning model.

REFERENCES

[1] Bayrak, E. A., Kırcı, P., & Ensari, T. (2019, April). Comparison of machine learning methods for breast cancer diagnosis. In *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)* (pp. 1-3). IEEE.

[2] Amrane, M., Oukid, S., Gagaoua, I., & Ensarİ, T. (2018, April). Breast cancer classification using machine learning. In *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)* (pp. 1-4). IEEE.

[3] Turgut, S., Dağtekin, M., & Ensari, T. (2018, April). Microarray breast cancer data classification using machine learning methods. In *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)* (pp. 1-3). IEEE.

[4] Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*, *83*, 1064-1069.

[5] Wang, Z., Li, M., Wang, H., Jiang, H., Yao, Y., Zhang, H., & Xin, J. (2019). Breast cancer detection using extreme learning machine based on feature fusion with CNN deep features. *IEEE Access*, *7*, 105146-105158.

[6] Samala, R. K., Chan, H. P., Hadjiiski, L., Helvie, M. A., Richter, C. D., & Cha, K. H. (2018). Breast cancer diagnosis in digital breast tomosynthesis: effects of training sample size on multi-stage transfer learning using deep neural nets. *IEEE Transactions on Medical Imaging*, *38*(3), 686-696.

[7] Rakhlin, A., Shvets, A., Iglovikov, V., & Kalinin, A. A. (2018, June). Deep convolutional neural networks for breast cancer histology image analysis. In *International Conference Image Analysis and Recognition* (pp. 737-744). Springer, Cham.

[8] Sun, D., Wang, M., & Li, A. (2018). A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE/ACM transactions on computational biology and bioinformatics*, *16*(3), 841-850.