Breast Cancer Detection using Machine Learning

Md. Sajeeb Molla(2018-2-60-045)[1], Aditta Ghosh(2018-2-60-048)[2]

[#]*Department of Computer Science and Engineering, East West University*
Aftabnagar, Dhaka-1212, Bangladesh
[1]2018-2-60-045@std.ewubd.edu
[3]2018-2-60-048@std.ewubd.edu

***Abstract*— Breast cancer is one of the most common and is causing a huge number of deaths in women. The high incidence and mortality of breast cancer is due to its considerably low accuracy of diagnosis. In this paper, we explore machine learning models that can be applied to help increase the accuracy of the diagnosis of breast cancer. The main problem of the project is to detect breast cancer based on a set of machine learning algorithms. We present a diagnosis model using both traditional and deep learning algorithms. Classical machine learning models including Random Forest, Logistic Regression, K Nearest Neighbor, Support Vector Machine and Neural Network models. We tested our models on the Breast Cancer Wisconsin dataset. Additionally, we applied feature selection and neural network models to improve the performance of the system. The paper demonstrates that machine learning models can be used for an automatic diagnosis for breast cancer.**

***Keywords*— Breast Cancer, Machine Learning, Logistic Regression, Neural Network, Random Forest, K-nearest neighbour, Support Vector Machine.**

I. Introduction

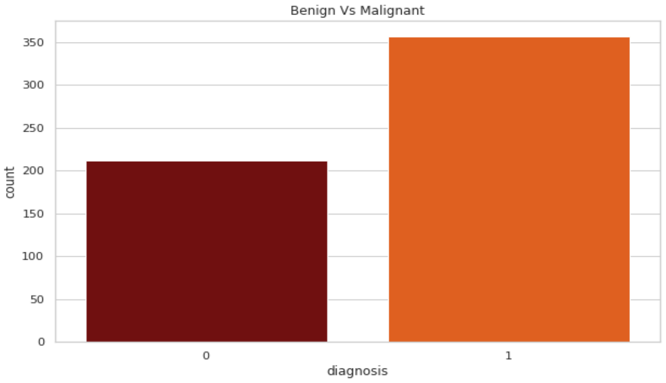
Breast cancer is one of the most common cancers in common and second most leading cause for women cancer deaths.Despite the lack of effective treatment, the low accuracy of diagnosis is also a major cause of the high incidence and mor- tality of breast cancer. According to UCHealth's report, only 78% of breast cancer can be accurately diagnosed by mammography. Many cases such as doctors' negli- gence or incompetence in addition to a mammography error may also result in a late diagnosis or misdiagnosis, which can be considered a cause of breast cancer death.In the long term, early-stage diagnosis could significantly increase the survival rate of breast cancer, therefore, it is important to improve the accuracy of breast cancer diagnosis.Machine learning has been applied in medical diagnosis in a large number of papers. In order to increase the accu- racy of breast cancer diagnosis, we aim to use machine learn- ing models and choose the model with higher performance.Breast Cancer Wisconsin is a widely used dataset provided by UC Irvine machine learning repository. In this paper, we will train our models using this dataset. We will use five traditional methods including Logistic Regression, Random Forest, Support Vector Machine, K-nearest Neighbour, and Neural Network.

II. Data-Preprocessing

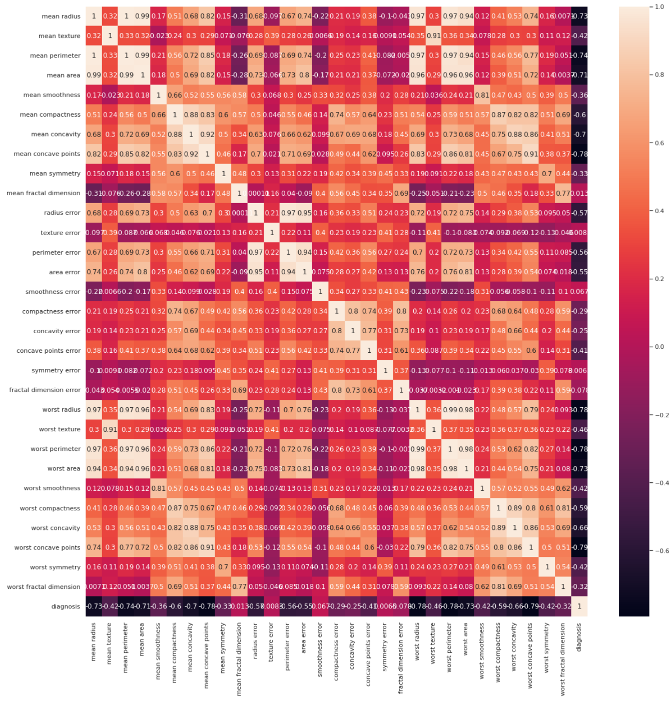
For pre-processing tha dataset we searched for null values and duplicate values. This dataset contains zero null values and zero duplicate values. All of these columns have no categorical attribute to encode. After analyzing the dataset we found some anomaly in the dataset. The dataset was imbalance. After realizing that we use some method to detect it and some other outlier value. To detect outliers, we use the IQR method. After finding outlier values we remove it.

II. Data Characteristics and Exploratory Data Analysis

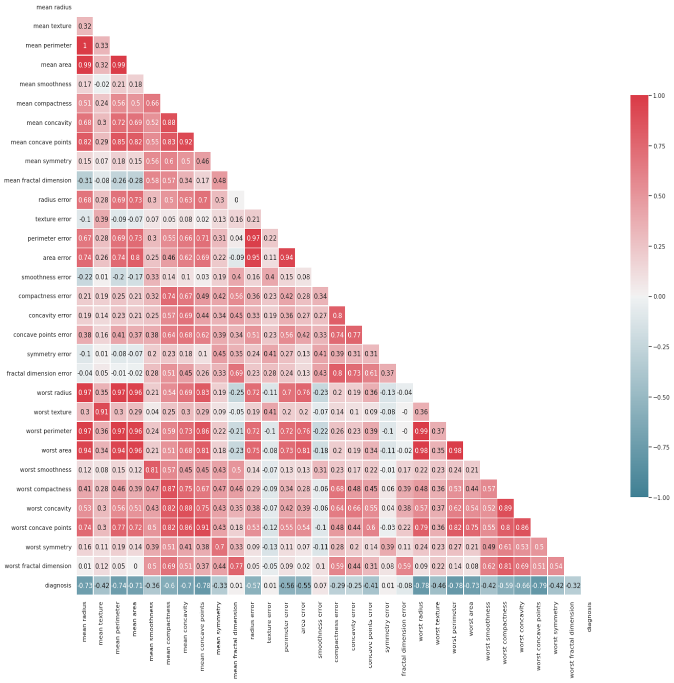
After reading the dataset "**Wisconsin dataset**" we calculated the summary where the dataset contains 569 rows and 31 columns. In this data set all of these columns are float but the decision making column is just integer.



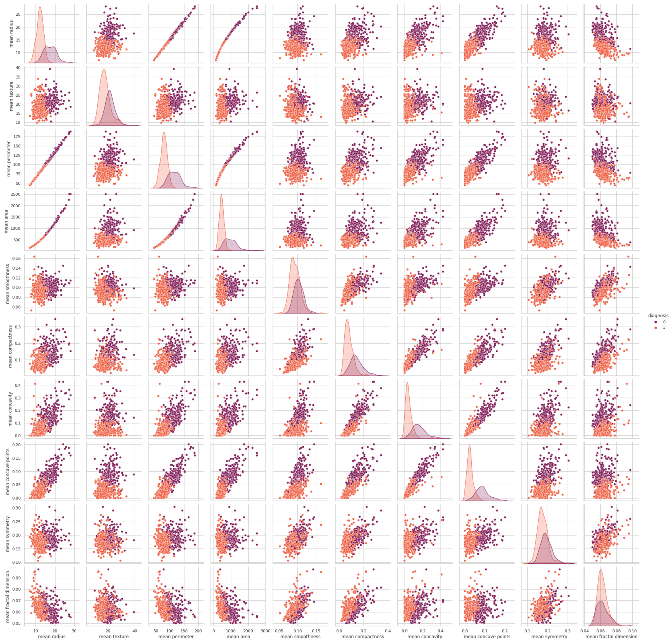
A tumor can be benign (not dangerous to health) or malignant (has the potential to be dangerous). Benign tumors are not considered cancerous: their cells aro close to normal in appearance, they grow slowly, and they do not invade nearby tissues or spread to other parts of the body. Malignant tumors are cancerous. In this graph we analyze 0 for benign and 1 for malignant.



At first sight, we see many positive correlations (red). However, this heatmap is messy. For visualization purposes, it would be better to group features that are highly correlated together. To do so, we will do hierarchical clustering.



A heatmap can display large amounts of data entirely because values are replaced by colors. This condensed color-coded format provides an easy-to-understand overview of data. It requires two dimensions and one measure. A second measure is optimal. The chart displays in a tabular format with color-coded tiles. The highest and lowest values show in each dimension column. The values in between are shown in a color gradient centered upon the average.



A pairplot visualizes given data to find the relationship between them where variables can be continuous or categorical. The pairplot function creates a grid of axes such that each variable in data will be shared in the y-axis across a single row and in the x-axis across a single column. That creates a plot as shown above.

IV. Algorithms

Logistic Regression is a supervised learning classification algorithm used to predict the probability of a target variable. It is one of the simplest ML algorithms that can be used for various classification problems such as spam dedication, cancer detection and so on. In our analysis the logistic regression accuracy score is approximately 92%.

Random Forest Classifier is a supervised machine learning algorithm that is used widely in classification and regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of this algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems. In our analysis the logistic regression accuracy score is approximately 94%.

Support Vector Machine (SVM) is a machine learning algorithm that analyzes data for classification and regression analysis. It is a supervised learning method that looks at data and sorts it into one of two categories. It outputs a map of the sorted data with the margins between the two as far apart as possible. It is widely used in text categorization, image classification, handwriting recognition and in the sciences. In our analysis the SVM accuracy score is approximately 96%.

K-Nearest Neighbor is a data classification algorithm that attempts to determine what group a data point is in by looking at the data points around it. It is basically a lazy learner algorithm because it does not generate a model of the data set beforehand. The only calculations it makes are when it is asked to poll the data points neighbors. This makes k-nn very easy to implement for data mining. In our analysis the K-Nearest Neighbor accuracy score is approximately 96%.

Neural networks, also known as artificial neural networks (ANNs) or simulated neural networks (SNNs), are a subset of machine learning and are at the heart of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another. In our analysis the Neural Network accuracy score is approximately 92%.

V. Algorithms

Algorithm	Accuracy Score	Precision Score	Recall Score
Logistic Regression	0.92	0.94	0.92
Random Forest	0.94	0.97	0.94
Support Vector Machine	0.96	0.98	0.95
K-nearest Neighbour	0.90	0.91	0.92
Neural Network	0.92	0.98	0.88

VI. Conclusions

This paper used two approaches: the regular ML approach and deep learning approach and the deep learning approach to predict breast cancer. In the DL approach, this paper proposes a neural network(ANN) model based on artificial neural networks. In the regular ML approach, RF, SVM, LR, KNN were compared with the optimized deep ANN. Three feature-selection methods: correlation matrix, univariate and accuracy-precision-recall were used to select the essential features from the database. The regular ML models and the optimized deep ANN are applied to selected features. The results show that the optimized deep ANN with selected features by univariate method has achieved the highest performance for cross-validation and testing results.

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2019,” CA: A Cancer Journal for Clinicians, vol. 69, no. 1, pp. 7–34, 2019.

[2] D. Dahiwade, G. Patle, and E. Meshram, “Designing disease prediction model using machine learning approach,” in Proceedings of the 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp. 1211–1215, IEEE, Erode, India, March 2019.

[3] N. F. Omran, S. F. Abd-el Ghany, H. Saleh, A. A. Ali, A. Gumaei, and M. Al-Rakhami, “Applying deep learning methods on time-series data for forecasting covid-19 in Egypt, Kuwait, and Saudi Arabia,” Complexity, vol. 2021, Article ID 6686745, 13 pages, 2021.

[1] Cruz JA, Wishart DS, Applications of Machine Learning in Cancer Prediction and Prognosis, Departments of Biological Science and Computing Science, University of Alberta Edmonton,AB, Canada.Vol.2, 2-21 (2006).

[2] Han J., Kamber M., Data Mining Concepts and Techniques. Morgan Kaufman Publishers, 2001.