International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 1, pp: 555-561 www.ijaem.net ISSN: 2395-5252

Predicting Breast Cancer Using Supervised and Deep Learning Algorithms

Rammohan Vadavalasa, Dr.Gali Nageswara rao

Date of Submission: 18-06-2020 Date of Acceptance: 03-07-2020

ABSTRACT:

The Changes in our environmental conditions accumulating health-related problems in day to day life. Across the globe, cancer is one of the leading cause for health-related deaths. The diagnosis of this problem at the earliest stages increases the survival rate.This helps the time of recovery and the cost associated with the treatment.

Diagnosis of breast cancer manually takes an extensive amount of time, there is a need to develop the automatic system for early detection of cancer. Machine learning and data mining technologies able to analyze large amounts of data and able to develop robust and predictive analytics. Effective and suitable machine learning algorithms need to be implemented in an adaptive platform, which can be applied for statistical computing and predictive analytics.

I. INTRODUCTION:

In today's society breast cancer is the most common cancer especially among middle-aged women around the world, according to global statistics it represents the highest cancer-related deaths on the globe. Worldwide, breast cancer is the eloquent health-related problem.

The chance of survival can increase significantly, only when it is predicted in the early stages. Proper classification of benign tumor can avoid patients experiencing additional treatments.Detecting the cancer in the early stages not only reduces the death rate but also increases the survival rate. In medical scenario breast cancer requires a medical diagnosis in the initial phase.Adopting computer-aided tools makes this process much simpler and quicker.In this paper, we classified the dataset into different types. Those are benign and malignant.

A mass of accumulation of an abnormal tissue is known as a tumor. In breast cancer tumors are classified into two types. 1) Benign, these are non-cancerous and 2) Malignant, these are cancerous. Benign tumors are not aggressive towards surrounding tissue and it lacks the ability to invade neighboring tissue. Malignant tumors are aggressive and cancerous because they can invade and damage the surrounding tissue. The patient undergo biopsy when a tumor is alleged to be malignant.

Approximately 10% of women affected with breast cancer at some stage in their life. Primarily genetic risk factors, menstrual periods, family history, aging, obesity and not having children increases the likelihood of developing breast cancer in females but the exact cause for breast cancer is still unknown.

We adopted the data mining methods for classifying and identifying understandable patterns from cancer data. Using these desirable patterns we can predict the cancer stage and its severity on a patient.

Data mining follows certain key steps:

1) Identifying the main goal

2) Data exploration: In this step involves controlling data accuracy, data sets size and patterns in the data.

3) Data preparation: It means cleaning and transforming the raw data for controlling missing and invalid values for further processing and robust analysis.

4) Modeling: Based on our goal for outcomes, a suitable algorithm should be selected for data analysis.

(5) Evaluation and Deployment:

Artificial Neural Networks, Decision trees and the K-Nearest Neighbor are the suitable algorithms for predicting patterns in breast cancer data sets.

Big data analytics plays a huge role here because in the future new datasets are getting added to the existing data. To handle such data an effective statistical models are selected.The correct classification of patients into malignant or benign groups are useful for the diagnosis of breast cancer. Unique advantage in critical features detection from complex breast cancer datasets, machine learning is widely recognized as the methodology of choice in pattern classification and forecast modeling.

For predicting breast cancer and its critical stage, supervised learning model Classification techniques K-NN (K-nearest neighbors) and SVM (Support Vector Machine) are used. In order to overcome many difficulties in feature-based approaches deep learning methods are becoming alternatives. Computer-aided diagnosis reduces the cost, complexity and increases the efficiency of the complete process because conventional classification methods rely on feature extraction methods, designed for a specific problem and enormous field-knowledge required for problem identification.

The data set used in this Paper contains 7,909 breast cancer histopathology images acquired from 82 patients[1]. The data set includes both benign and malignant images. substantial growth in the tumor size decreases the curability percentage.
In this paper, the model predicts the tumor severity stage and how it is useful for increasing the survival rate.

II. RELATED WORK:

About one in eight women are diagnosed with breast cancer during their lifetime. The most common type of breast cancer is Ductal Carcinoma[2]. If it is detected in its early stages, there is a good chance of recovery.

Breast cancer is identified with different stages. It contains 4 types of stages depending upon tumor size. Breast cancer stages are:

Stage 0: The tumor measures 0 centimeters (cm) - Benign (Non-Cancerous).
Stage 1: The tumor measures up to 2 centimeters (cm) – Malignant (Cancerous). Stage 2: The tumor measures 2 to 5 centimeters (cm) – Malignant (Cancerous). Stage 3: The tumor measures 5 to 9 centimeters (cm) – Malignant (Cancerous). Stage 4: The tumor measures above 9 centimeters (cm) – Malignant (Cancerous).

The above stages can be identified by a blood sample test. Stage 4 is not curable.
In this paper we are using numerical data from blood sample test results.

Machine learning, is a cutting-edge field in computer science that seeks to get computers to carry out tasks without being explicitly programmed to carry out a given task[3]. Machine learning uses many techniques to create algorithms, in order to learn and make predictions from data sets. In data mining machine learning is used to discover patterns and models in data sets, where relationships are previously unknown.

These analytical models allow researchers, data scientists, engineers, and analysts to "produce reliable, repeatable decisions and results" and uncover "hidden insights" through learning from historical relationships and trends in the data.

Machine Learning can be done by performing class of a task, performance measurement and experience. Any Machine Learning Algorithm can follow some steps to develop the model and predict the data.

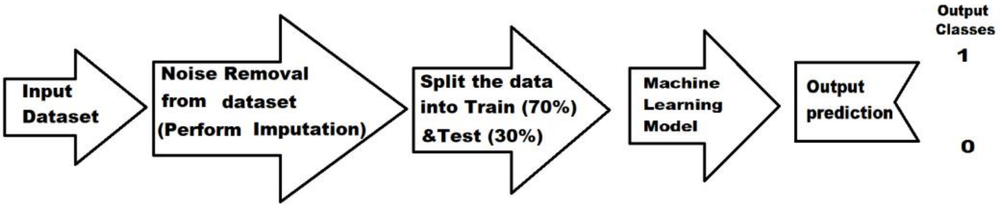
For application development FLASK application is used and written in python framework. Falsk works together with Werkzeug WSGI and jinja2.
For python web application development, WSGI (web server gateway interface) has been adopted as a standard tool. Werkzeug is a WSGI toolkit and it used for requests, response objects, and other utility functions implementation.

III. RESULTS:

1) Problem Framing: It‘s values should be predicted from the data set and how we approach from problem statement to solution.

2) Data collection: Collecting related data set from related data sources and using python transferring CSV file into the data frame. Later removing missing values and tuning data for model preparation.
3) later we evaluate which features are useful for predicting the end solution and which features are not suitable for the machine learning process because feature selection directly influences target end results. Based on our features and their binomial format, we have to select a suitable machine learning algorithm for data processing.

Classification algorithms and regression techniques are used to develop predictive models in supervised learning. In classification algorithms, inputs are divided into classes and these inputs are related to one of these classes. This is a typical supervised learning process as shown in Fgure3.1.


Input Dataset
Noise Removal from dataset (Perform Imputation)
Split the data into Train (70%) &Test (30%)
Machine Learning Model
Output prediction
Output Classes
1
0

3.1 Architecture for supervised learning

K-nearest neighbors is used for pattern recognition and statistical estimation, and it uses distance functions to classify new cases based on available cases.
K-NN is a non-parametric supervised learning technique in which we try to classify the data point to a given category with the help of a training set [4].

Based on majority of its neighbor votes new case is classified, the chosen case is most common among other k nearest neighbors. Predictions are made for a new instance (x) by searching through the entire training set for the K most similar cases (neighbors) and summarizing the output variable for those K cases

Euclidean $$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

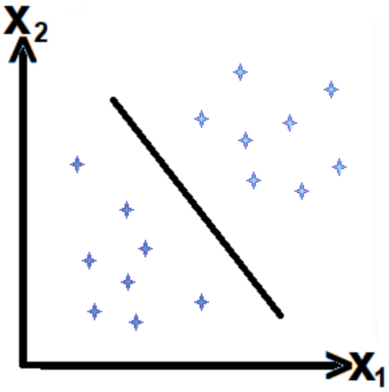
Manhattan $$\sqrt{\sum_{i=1}^{k}|x_i - y_i|}$$

Minkowski $$\left[\sum_{i=1}^{k}(|x_i - y_i|)^q\right]^{1/q}$$

3.2 Distance metrics

The distance measure types are Euclidean, Manhatten, Minkowski, valid for continuous variables as shown Figure3.2. Hamming distance is used for categorical variables for nearest neighbors. For example, Euclidean distance is used when two points in a plane are A(x0,y0) and B(x1,y1)

The **support vector machine** is a supervised machine learning model used for classification and regression problems but it is most commonly used in classification. SVMs will find the best suitable hyperplane that divides a dataset into two classes.



3.3 support vector machine Hyperplane Graph

A line that classifies and linearly separates a set of data is called a hyperplane as shown in Figure 3.3. Being data points away from hyperplane, provides better accuracy, when we add new data to the model. It will land on the hyperplane, where it's features suit.

Margin known as the distance between the nearest data point and hyperplane. Choosing hyperplane that is the greatest possible margin between training set and hyperplane, because this margin will influence the new data being correctly classified.
Here linear kernel is used for separating data between classifiers.

Hyperplane equation is WTX=0; which is similar to line equation Y = ax +b. W and X represented as vectors and WTX represents the dot product of vectors.

The hyperplane of SVM is built on mathematical equations. The equation of hyperplane is WTX=0 which is similar to the line equation y= ax + b. Here W and X represent vectors where the vector W is always normal to the hyperplane.WTX represents the dot product of vectors. In the SVM model each data item is presented in an n-dimensional space, where n represents the number of features. In this n-dimensional space each feature is represented as the value of a particular coordinate.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

3.4 Confusion Matrix

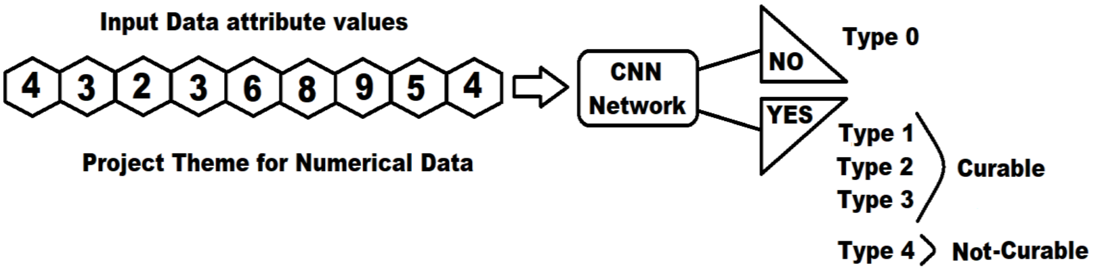
3.5

A confusion matrix as shown in Figure.3.4, predicts the N number of classes and it is a N X N matrix. If N = 2 then we get 2 X 2 matrix

The total number of prediction proportion is called accuracy, the proportion of corrected positive cases is called positive predictive value or precision, negative cases are called negative predictive value, correctly identified positive cases are called sensitivity or recall and finally correctly identified negative cases are called specificity.

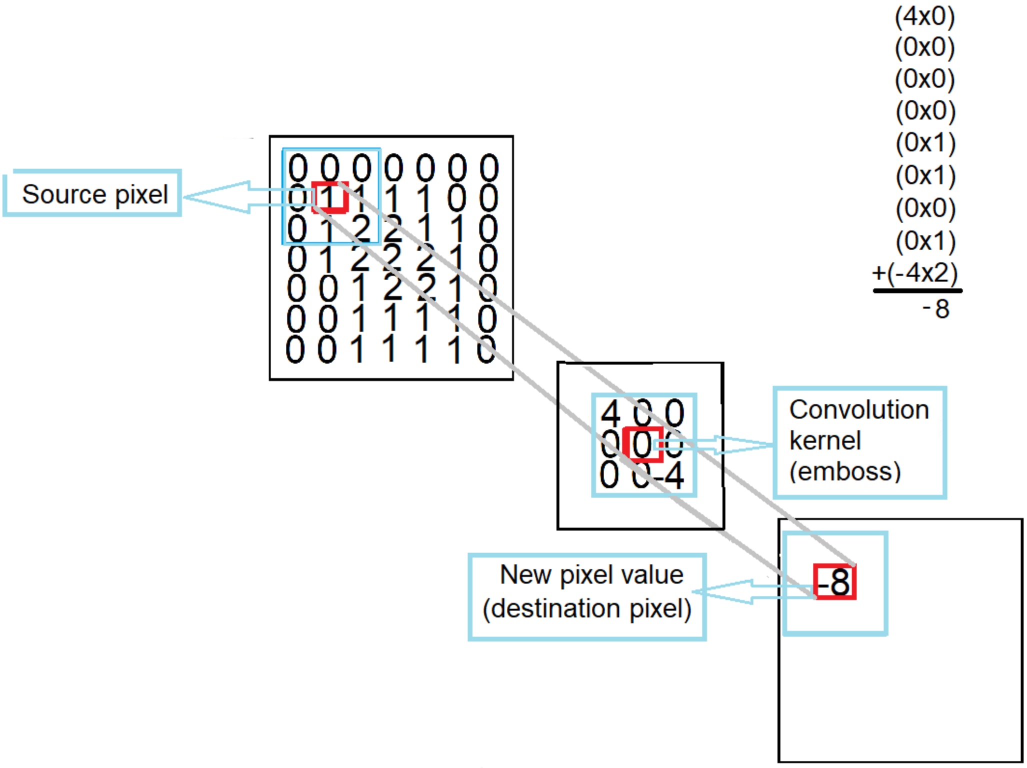
We created a web application using CSS, javascript and python micro web framework called Flask.

Here, we perform a model for predicting the survival rate at an early stage but increasing the tumor size reduces cancer curability. We take training data and train the data set using KNN and SVM models.



3.5 Architecture for Conventional Neural Network

A deep learning algorithm is a convolution neural network as shown in Figure 3.5,here, we train data for predicting cancer stage using Conventional Neural Network algorithm. It can take input as an image or numerical data and assign importance, based on various aspects, to be able to recognize one from the other.



3.6 Convolution Operation in Conventional Neural Network

The architecture and inspiration of a convolution neural network are taken from the connectivity pattern of neurons and the visual cortex in the human brain.

Using relevant filters convolution neural network successfully captures the Spatial and Temporal dependencies for every given input. Convolution neural network can reduce the input data and process it for good prediction, without losing its original features. Kernal is the first part of the convolution Layer in convolution operation as shown in Figure 3.6.

Later valid padding can reduce the dimensionality of the data, as compared to the input data, otherwise remains the same as earlier and spatial size reduction is done by the Pooling layer.

After completion of the above process, the model can successfully understand the main features and can flatten the final output. Flattened output is again fed into back-propagation for classifying low-level features using the softmax Classification technique.

TEST CASES:

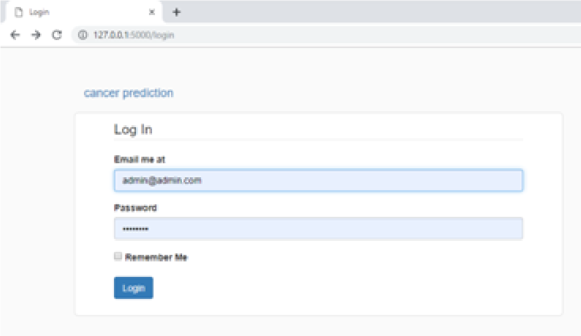
S. No	Requirement ID	Test Case ID	Type	Test Case Description	Test Data	Expect Value	Actual Value	Result	Priority ID
1	R1	101	+VE	Fill the login form with correct Email & Password	Email: admin@admin.com Password: password	Login to admin page	Login to admin page	Login	10

2	R2	102	+ve	Take the data set and check all the fields are there or no	c.t='2', cs='3', csh='2', ma='4', bn='4, bc='2', nn='3', m='2'	Data set will be sent to processing and it is redirected to next page	Non – Cancerous	Cancerous stage predicted to 2	10
3	R3	103	+Ve	Fill all the fields and submit	c.t='9', cs='3', csh='8', ma='6', bn='4, bc='4', nn='7', m='6'	data set will be sent to processing and it is redirected to next page	Cancerous	Cancerous stage predicted to 4	9

Table 1: Test Cases

Testing is conducted using different test cases as shown in Table 1

INPUTS & OUTPUT:



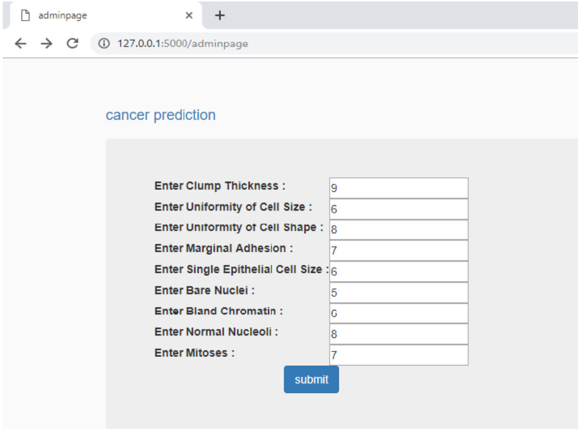
Log In

Email me at: admin@admin.com

Password: ********

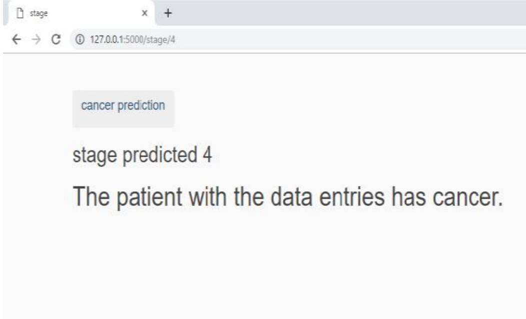
□ Remember Me

screenshots 1: Application Log In



Enter Clump Thickness : 9
Enter Uniformity of Cell Size : 6
Enter Uniformity of Cell Shape : 8
Enter Marginal Adhesion : 7
Enter Single Epithelial Cell Size : 6
Enter Bare Nuclei : 5
Enter Bland Chromatin : 6
Enter Normal Nucleoli : 8
Enter Mitoses : 7

screenshots 2: submitting blood sample results



stage predicted 4

The patient with the data entries has cancer.

screenshots 3:Application output

We have to enter blood sample test results in the web application, once we entered and click on submit, based on entered values KNN and SVM predicts entered numerical value has cancer or not and if it contains cancer, it will display including the cancer stage using Conventional Neural Network as shown in screenshots 1,2,3.

We can predict the cancer stage based on its numerical value index, if the tumor size is more than 9 centimeters, it is classified as a stage 4 and not curable. We displayed the cancer stage on the monitor using a flask web application as shown in screenshots 1,2,3.

IV. CONCLUSION:

This study attempts to detect breast cancer in early-stage using a machine learning algorithm. In order to overcome the limitations of existing privacy models and identify the early diagnosis of breast cancer, this model is recommended. Big

Data, data mining and machine learning technologies are used in this study.

Feature selection and dimensionality reduction and integration of multidimensional data techniques helped for improving better model building and predicting better outcome.

Study and experiment are conducted using different Python libraries. Finally, the proposed model is useful for automatic diagnosis of breast cancer and predicting its severity.

REFERENCES:

[1]. Breast Cancer data set (Numerical) UCI repository http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/

[2]. „illness and conditions“, „breast cancer female“ - NHS business serive authority United Kingdom

[3]. Ashim Saha, Nirmalya Kar, Suman Deb “Advances in Computational Intelligence, Security and Internet of Things: Second International Conference“, ICCISIoT 2019, Agartala, December 13–14, 2019, Proceedings Springer Nature

[4]. Deepanshu Bhalla „K Nearest Neighbor article “, „listen data“

[5]. Karleigh Moore, Asrafe MOSSUS, Simone Dhu „machine learning“,“brilliant.org“

[6]. José Rouco,Paulo Aguiar, Catarina Eloy, António Polónia, Aurélio Campilho „Classification of breast cancer histology images using Convolutional Neural Networks“ Research Article ; Published: June 1, 2017

[7]. B. Zheng, S. W. Yoon, and S. S. Lam, ” *Expert Systems with Applications*, vol. 41, no. 4, pp. 1476–1482, 2014. “Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms,

[8]. H. Dhahri and A. M. Alimi, “Opposition-based particle swarm optimization for the design of beta basis function neural network,Barcelona, Spain, July 2010” in *Proceedings of the .e 2010 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8,

Rammohan Vadavalasa "Predicting Breast Cancer Using Supervised and Deep Learning Algorithms." *International Journal of Advances in Engineering and Management (IJAEM)*, 2(1), 2020, pp. 555-561.

International Journal of Advances in Engineering and Management
ISSN: 2395-5252

IJAEM

Volume: 02 Issue: 01 DOI: 10.35629/5252

www.ijaem.net

Email id: ijaem.paper@gmail.com