A Predictive Model of Breast Cancer Survivability using Support Vector Machine

Shilpa Balan, Nikita Dhanraj Marathe, Ritika Joshi

Big Data in Precision Health 2019, Stanford University, CA

Introduction

- In the United States, approximately one in eight women has a risk of developing breast cancer [1].
- An analysis from the American Cancer Society has shown that the survival rate is 88% after 5 years of diagnosis and 80% after 10 years of diagnosis [1].
- As per the 2019 report published by American Cancer Society, about 606,880 US citizens are expected to die of cancer in 2019, which translates to about 1,660 deaths per day [2].
- This study is an illustration of a predictive model of breast cancer, which can help clinicians to perform early stage diagnoses on individuals who are at a higher risk.
- Our key focus in this study is to build a predictive model to help predict whether the breast cancer cells are benign or malignant.
- Survivability rate of patients with breast cancer can be increased using the model built in this study.

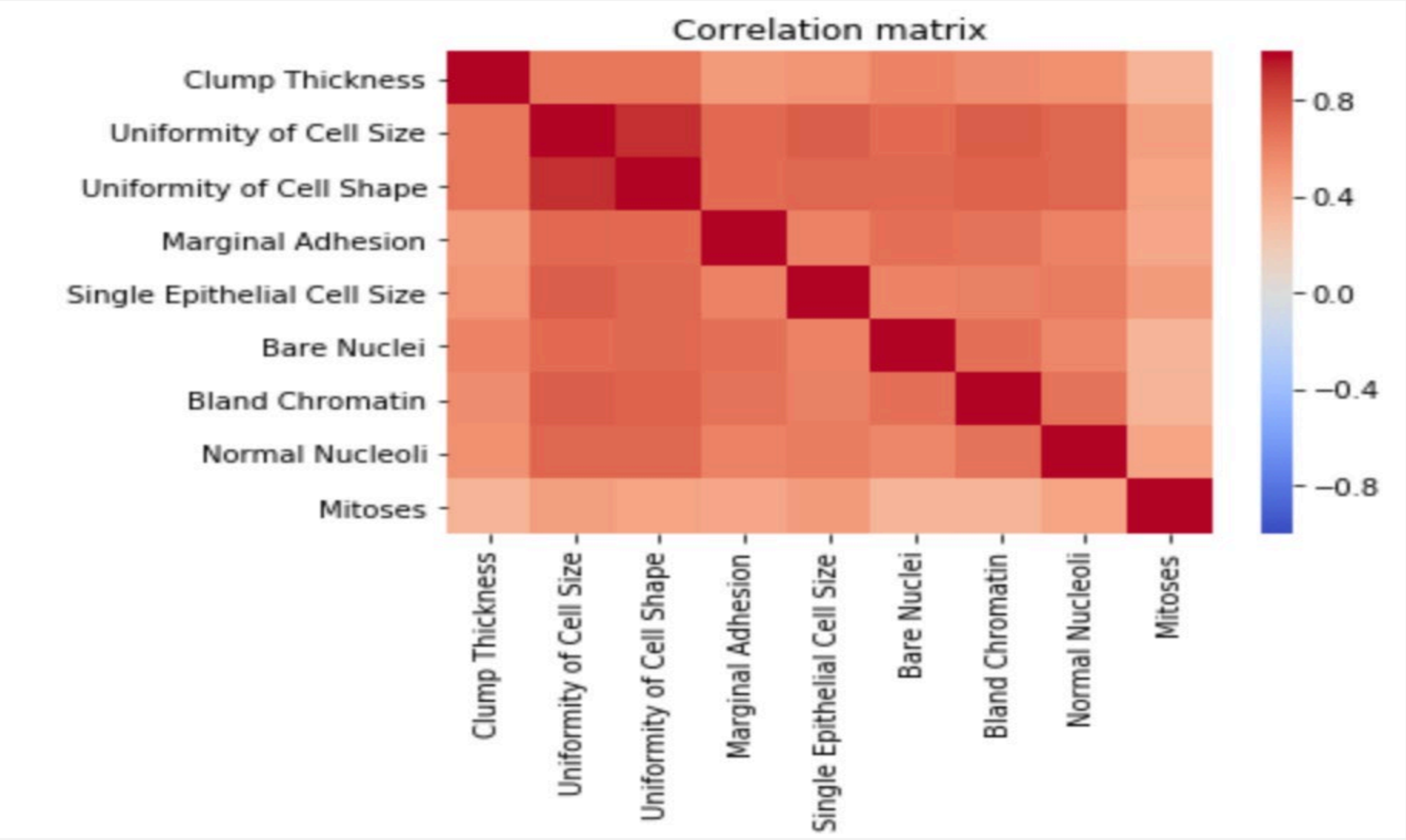
Background

- The second leading cause of death among women is breast cancer, as it comes directly after lung cancer [6].
- Data mining approaches in medical domains is increasing rapidly due to the improvement effectiveness of these approaches to classification and prediction systems, especially in helping medical practitioners in their decision making.
- In addition to its importance in finding ways to improve patient outcomes, reduce the cost of medicine, and help in enhancing clinical studies.
- Although there was a great deal of public education and scientific research, Breast cancer considered the most common invasive cancer in women, with more than one million cases and nearly 600,000 deaths occurring worldwide annually [7].

Methodology

- The dataset used in this study was created using samples collected from Wisconsin hospital over a period of Jan'89 to Nov'91 [3].
- The data used in our study is a breast cancer data set that is available for public use from the UCI machine learning repository [3].
- Fields that impact the breast cancer prediction are *Marginal Adhesion, Single Epithelial Cell Size, Bland Chromatin, Normal Nucleoli, Mitoses, Clump Thickness and Bare Nuclei*.
- Support Vector Machine (SVM) algorithm is used to build the predictive model in our study, as it analyses the training data and results it into an optimal hyperplane (i.e. a line dividing a plane in two parts where in each class lay in either side) which categorizes new examples [4].
- For linear kernel the equation for SVM prediction is calculated as: f(x) = B(0) + sum(ai * (x,xi)), where x is the input, support vector is x(i), B(0) and ai (for each input) are the coefficients [5].

Analysis



	PRECISION	RECALL	F-MEASURE
BENIGN	94%	99%	96%
MALIGNANT	98%	88%	93%

```
import pandas as pd
import numpy as np
from sklearn import svm
import seaborn as sns
import sklearn.model_selection as ms
from sklearn.model_selection import train_test_split

Read CSV data file and check for missing values
Convert Data type to numeric for Mitoses, Clump Thickness, and Bare Nuclei
Check correlation between selected features to avoid bias
    correlation_matrix = df1.corr()
    sns.heatmap(correlation_matrix, vmax=1.0, vmin=-1.0, cmap='coolwarm')
    Remove highly correlated variables
Split 60% data for testing and 40% data for training
Set independent variables as Marginal Adhesion, Single Epithelial Cell Size, Bland
Chromatin, Normal Nucleoli, Mitoses, Clump Thickness and Bare Nuclei
Set dependent variable as malignant/non-malignant
Apply SVM (Support Vector Machine) algorithm to train and test data
    clf = svm.SVC(kernel='linear')
    metrics = sklm.precision_recall_fscore_support(actual, predicted)
    conf = sklm.confusion_matrix(actual, predicted)
Measure the prediction accuracy (95%)
    sklm.accuracy_score(actual, predicted)
```

Pseudocode for Breast Cancer Prediction

Results

- The two features Uniformity of Cell Size and Cell Shape were not useful in the prediction for breast cancer diagnosis due to its high correlation.
- Recall can be thought as of a model's ability to find all the data points of interest in a dataset [8].
- Precision is the ability of a classification model to identify only the relevant data points [8].
- We found the SVM model built in this study to have a high recall and precision, making our model more accurate and relevant, as the model has the ability to extract relevant malignant and benign cases.
- From the SVM model built in this study, our prediction accuracy for breast cancer diagnosis was found to be 95%.

Conclusion & Future Research

- The discovery of the survival rate or survivability of a certain disease is possible by extracting the knowledge from the data related to that disease.
- This study has outlined the issues, algorithms, and techniques for the problem of breast cancer survivability prediction
- This study clearly shows that the results are promising for the application of the data mining methods into the survivability prediction problem in medical databases.
- For future research, we would like to examine more current breast cancer data sets and apply our model to it.
- For future research, we plan to also implement the ANN (Artificial Neural Network) model and compare results with the SVM model built in this study.

References

[1] American Cancer Society. Breast Cancer Facts & Figures 2005-2006. Atlanta: American Cancer Society, Inc., Retrieved from http://www.cancer.org/

[2] Cancer Facts & Figures 2019. American Cancer Society, 2019. Web. 19 Apr 2019. Retrieved from http://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2019/cancer-facts-and-figures-2019.pdf

[3] UC Irvine Machine Learning Repository (2019). Retrieved Apr 25, 2019, from https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data

[4] Tutorials, O. (2014). Introduction to Support Vector Machines. Retrieved May 5, 2019, from https://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html

[5] Patel, S. (2017). Chapter 2 : SVM (Support Vector Machine) — Theory – Machine Learning 101. Retrieved May 5, 2019, from http://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72

[6] U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2012.

[7] Lyon IAfRoC: World Cancer Report. *International Agency for Research on Cancer Press* 2003:188-193.

[8] Koehrsen, W. (2018). Retrieved May 2, 2019, from https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c