Topics in Data-Driven Analysis and Computation

Statistical Learning in Biological and Information Systems

EECS E6690

Final Project Report

By

Ananye Pandey
UNI: ap3885

Ruturaj Rajendra Nene
UNI: rn2494

Raksha Nandanahosur Ramesh
UNI: rn2486

COLUMBIA UNIVERSITY

2020

Abstract

Breast Cancer is the leading cause of cancer death in women worldwide. It is also difficult to diagnose. Moreover, misdiagnosis can lead to unnecessary invasive biopsies and delayed diagnosis can be fatal. Machine learning can therefore play an important role in increasing the accuracy of cancer diagnosis. The primary goal of the study is to develop a prediction model for breast cancer diagnosis based on features obtained from an FNA test. To improve the classification accuracies, data is pre-processed and different feature selection methods are tested and a comparative analysis is performed.

Table of Contents

Introduction... 1

Section 1: Objective & Problem Formulation 3

Section 2: Description of Dataset 3

Section 3: Summary of Paper 4

3.1 Methodology... 5

3.1.1 Training and Testing Details... 6

3.2 Summary of Results from Paper.. 6

Section 4: Implementation.. 7

Section 5: Results and Discussion 13

References or Bibliography 17

Introduction

Breast cancer is one of the leading causes of women's deaths in the world today regardless of race and ethnicity [1]. According to statistics, a woman living in the US has a 12.4 percent lifetime risk of being diagnosed with breast cancer, this translates to 1 in every 8 women today [2]. The causes of breast cancer include family history and genetic causes. The presence of BRCA1 and BRCA2 genes can increase the risk of breast cancer by 40-80% when inherited [3]. Hormonal changes and lifestyle changes are other causes that contribute to the risk factor.

Breast cancer tumors can be broadly categorized into - benign (non-cancerous), premalignant and malignant tumors. Malignant tumors are a result of abnormal cell growth which can be life threatening as the disease progresses. It is typically diagnosed through a screening examination with a mammogram, MRI or Ultrasound. A biopsy of the breast tissue is required to understand the extent of spread if cancer is suspected. This is done with a breast fine needle aspiration (FNA) which is used to extract fluid from a breast lesion. The FNA test can help the health care practitioner to understand the lesion and plan future treatment in case of a positive diagnosis. However, if cancer goes undetected, it can metastasize by invading nearby healthy cells and travelling through the circulatory system. Thus, early diagnosis and intervention is pivotal to prevent disease progression.

Machine Learning for Breast Cancer Detection

With the advancements in imaging technology, along with the availability of data, machine learning techniques can play a vital role in helping health care professionals with detection and diagnosis of cancer. According to statistics, one in every 10 cancers is misclassified as non-cancerous [4]. Additionally, the likelihood of false positives in mammograms are high, causing patients to be subjected to unnecessary biopsies and invasive intervention. The main challenge for Machine Learning algorithms therefore is to reduce the false positives and false negatives. In recent years, the accuracy of cancer prediction models has increased by 15-25% [5].

Various studies conducted on the WBCD dataset have been able to apply common machine learning techniques to address cancer detection and diagnosis. Bazazeh and Shubair [6] use SVM, random forest and Bayesian networks. According to their results, SVM outperformed other classifiers with respect to accuracy, precision and specificity. Random forest had the highest probability of correctly classifying tumors. Chaurasia , [7] use Naïve Bayes, RBF and J48 variant of decision trees with 10-fold cross validation and compare model performance. Saxena S[8] implements MLP, RBFNN and other ANNs for classification and presents a comprehensive review of these techniques to diagnose breast cancer.

Section 1: Objective & Problem Formulation

- Classification between Malignant and Benign tumor for Breast Cancer Detection
- Feature Engineering and testing different classification methods to improve accuracy
- Comparison of results with reference paper

Section 2: Description of Dataset

The dataset used for the analysis is the **Wisconsin Breast Cancer Diagnositc Dataset** obtained from UCI ML repository.

The dataset consists of 569 instances of 32 attributes – diagnosis (dependent variable), IDs of patients and 30 real value attributes. It is a multivariate, classification problem.

The cytological features are extracted from digital scans obtained from biopsy of the breast mass carried out by FNA test. The ten features are -

- Radius of cell nucleus
- Perimeter
- Area
- Compactness - a measure of compactness derived from the perimeter and area of cell
- Smoothness of nuclear contour
- Concavity - measure of number and severity of indentations in cell nucleus
- Concave points - measures the number of contour concavities
- Symmetry - The difference in lengths between major axis (longest chord) and chord perpendicular to the longest chord.
- Fractal dimension
- Texture - variance of gray scale intensities in pixels

The features are modeled such that higher values are associated with malignancy. For example, a higher value of fractal dimension translates to an irregular contour which might indicate a higher probability of malignancy. The above ten features are calculated for each cell in the sample and summarized by computing the mean, extreme value, Standard error, creating thirty attributes in total.

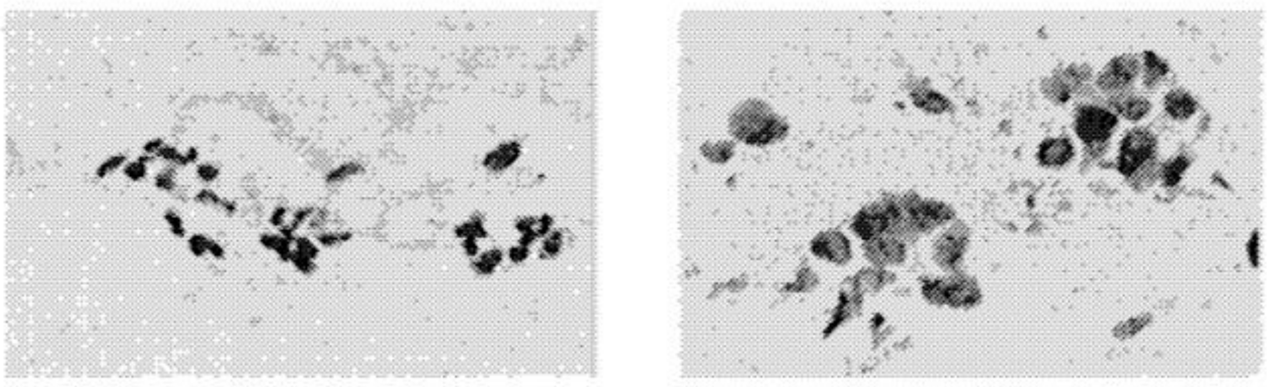


Figure 1 : Digitized Images of FNA sample, Benign (*left)* Malignant (*right*) [ref no]

Section 3: Summary of Paper

Reference Paper: *Abien Fred M. Agarap. 2018. On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset. In ICMLSC 2018: ICMLSC 2018, The 2nd International Conference on Machine Learning and Soft Computing, February 2–4, 2018, Phu Quoc Island, Viet Nam.*

The paper presents a comparison between six machine learning algorithms for the classification task namely – GRU-SVM, Linear Regression, Multilayer perceptron, Softmax regression, Nearest neighbor, Support Vector machines and evaluates these models based on test accuracy, sensitivity and specificity. The dataset used in the paper is the above described Wisconsin Diagnostic Breast Cancer Dataset.

3.1 Methodology

Pre-Processing: The data is first standardized – zero-centered and normalized by standard deviation before classification.

Classification Techniques:

1) GRU-SVM: This combines the gated recurrent unit from RNNs with SVM. The state value of the GRU unit is used as the predictor variable x in the L2-SVM predictor function. The hinge-loss function is optimized with Adam. The scores obtained then correspond to the probability of the predicted classes.

2) Linear Regression: Since the problem is a binary classification, the paper applies a threshold on the output of the linear regression to classify into the two classes. The loss is measured with Mean squared error and minimized with Stochastic gradient descent.

3) MLP: Multilayer Perceptrons are feed forward networks with multiple hidden layers that have activation functions to model non-linearities of the data. The weights of the network are trained via backpropagation in order to minimize the cross-entropy loss. The paper uses ReLu activation with three hidden layers and 500 neurons in each layer.

4) Nearest Neighbor: The nearest neighbor classifier assigns a point to the class of its closest neighbor. The paper uses L1 and L2 norms to measure the distance.

5) Softmax Regression: This method outputs a probability distribution for the classes with the softmax function.

6) Support Vector Machine: The objective of SVM is to compute the optimal hyperplane that best separates the two classes. L2-SVM contains the L2 regularization parameter to make SVM

less susceptible to outliers i.e L2 SVM uses squared sum of the slack variables in the objective function.

3.1.1 Training and Testing Details

The paper uses a 70/30 train test split respectively. For evaluation, FPR, FNR, TPR and TNR are computed.

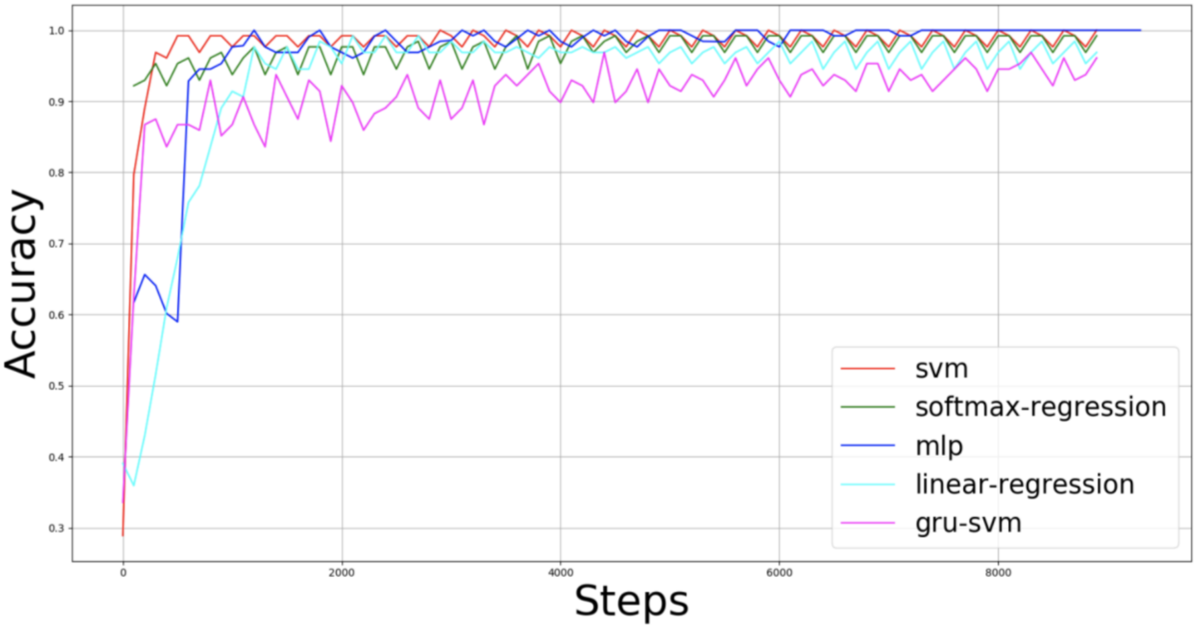


Figure 2: Training Accuracies of the various classification algorithms[ref no]

3.2 Summary of Results from Paper

Metric/Model	GRU-SVM	Linear Regression	MLP	L1-NN	L2-NN	Softmax Regression	SVM
Accuracy	93.75%	96.09%	99.03%	93.56%	94.73%	97.65%	96.09%
FPR	16.67%	10.20%	1.26%	6.25%	9.38%	5.769%	6.38%
FNR	0	0	0.78%	6.54%	2.80%	0	2.47%

TPR	100%	100%	99.21%	93.45%	97.19%	100%	97.53%
TNR	83.33%	89.80%	98.73%	93.75%	90.06%	94.23%	93.61%

Figure x presents the training accuracies of all the algorithms tested.

GRU-SVM had an average training accuracy of 90.68%, Linear Regression had an average training accuracy of 92.89%, MLP and softmax regression had an average training accuracy of 96.92% and 97.36% respectively. L2-SVM had a training accuracy of 97.73%. Figure x presents the training accuracies of all the algorithms tested.

Comparing the testing accuracies from the above table, GRU-SVM does not perform as well as the other classifiers. Since the WDBC dataset is linearly separable, the non-linearities GRU cell introduces because of its gating mechanism, may be the reason for the model's poor performance. Also, RNNs are sensitive to weight initialization. Therefore, for the project, we do not implement the GRU-SVM classifier. Instead we test out a few more classification methods and analyze the best combinations of feature selection and classification methods for such data. The paper also does not carry out k-fold cross validation, which we implement to evaluate each of our models.

The following sections describe our implementation.

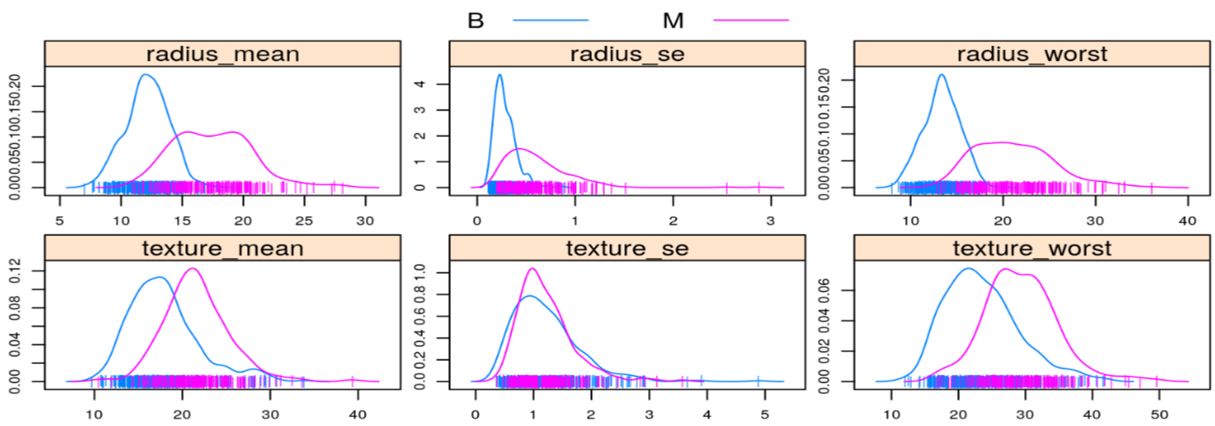
Section 4: Implementation

Exploratory Data Analysis and Feature engineering:

We started the data exploration by finding out what is the summary of the data.

The next key observation about the dataset was the distribution of the diagnosis (target) variable of the dataset. There are 212 Malignant and 357 Benign cases in the dataset. In terms of percentage, the malignant tumor is 37.3%. The percent is unusually large; the dataset does not represent a typical medical analysis distribution where we will have a considerable large number of cases that represent negative vs. a small number of cases that represent positives (malignant) tumors.

We further analyzed the features and tried to understand which features have larger predictive value and which does not bring considerable predictive value if we want to create a model that allows us to guess if a tumor is benign or malignant. For this we plotted the density plot, to represent both the values density and the degree of separation of the two sets of values, on each feature direction. We observed there is no perfect separation between any of the features; but there were fairly good separations for concave.points_worst, concavity_worst, perimeter_worst, area_mean, perimeter_mean.



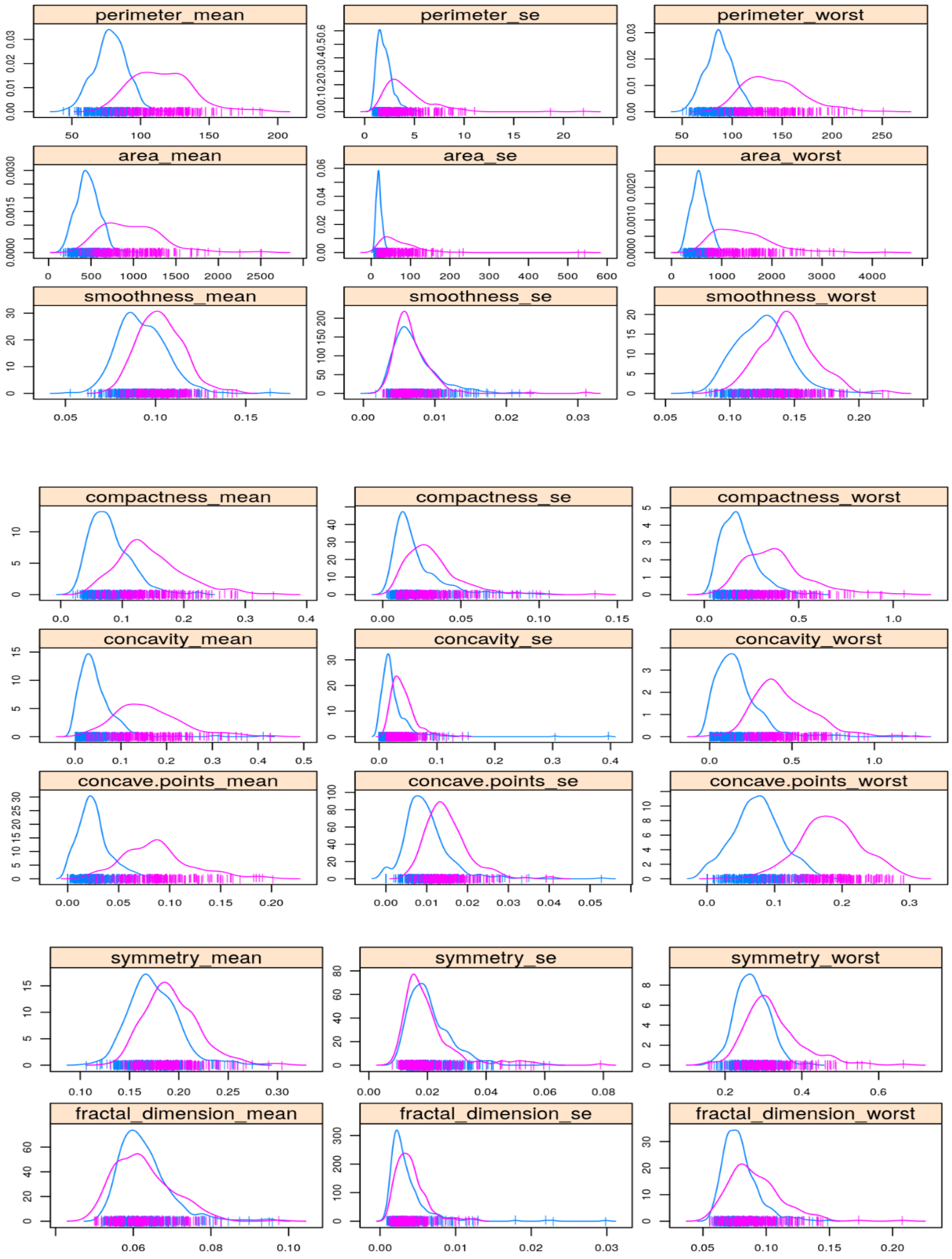


Figure 3 : Density plots for each feature separated by class

As there is no separable boundary between two features in the dataset, we decided to transform the dataset into a new vector space. To do this we tried two different procedures - Principle component analysis (PCA) and Linear Discriminant Analysis (LDA). These two methods will not only project data into new dimensions but also reduce the dimensionality.

Dimensionality Reduction

1) **Principal Component Analysis**

The principal component analysis is a feature extraction method that transforms the feature space into a linear combination of the original set of features. The main goal of PCA is to find the component axes that maximizes the variance in our data.

Although PCA makes sure the variables are uncorrelated and independent, it compromises on the interpretability of these features.

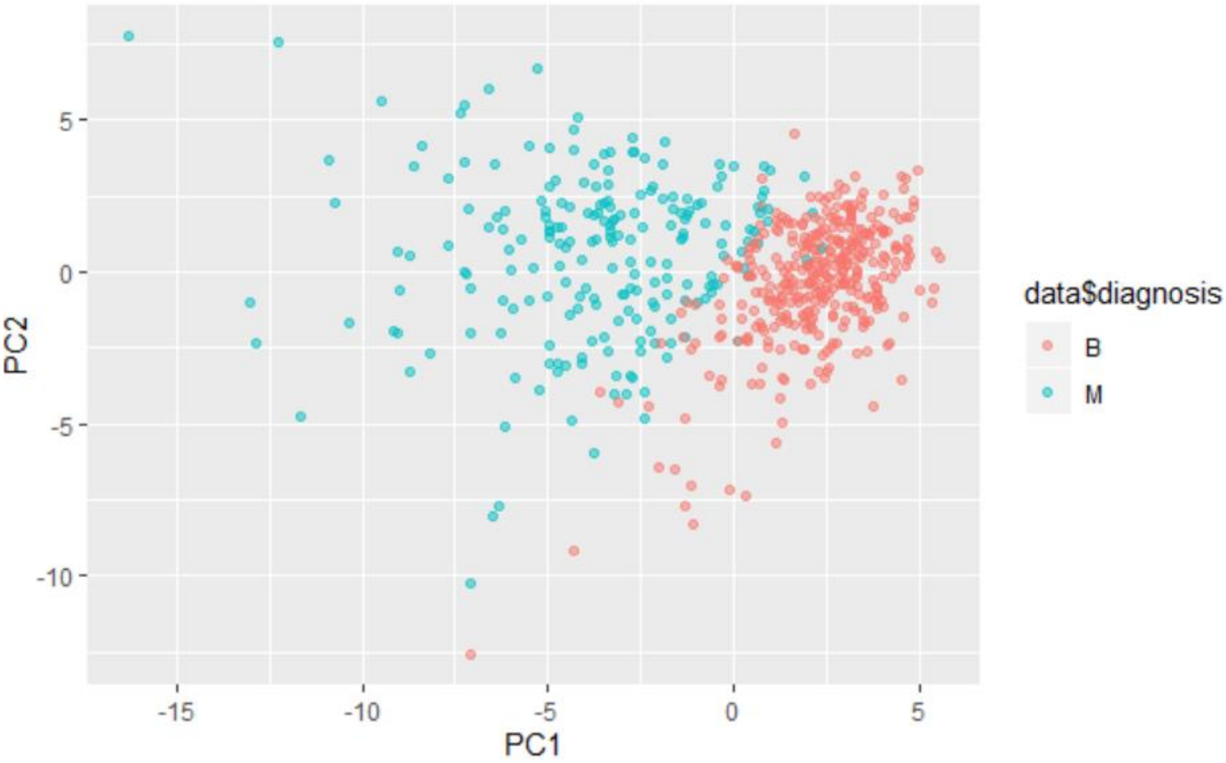


Figure 4 : Scatterplot

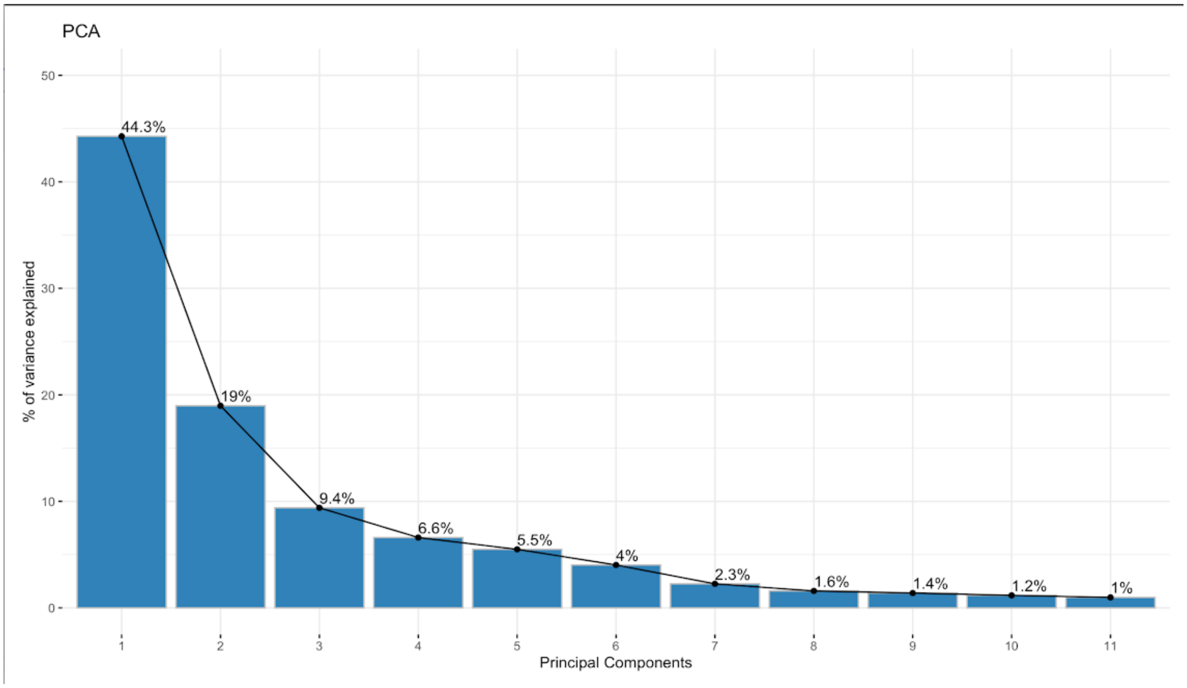


Figure 5: Scree-plot that explains proportion of variances of each PC

PCA works effectively when we retain components that explain up to 99% of the variance.

2) **Linear Discriminant Analysis**

The goal of LDA is to project the features onto a new feature space that ensures maximum class separability. Since the data is normally distributed and both classes have equal covariance matrices, LDA is an optimal choice for reducing the dimensionality in the dataset.

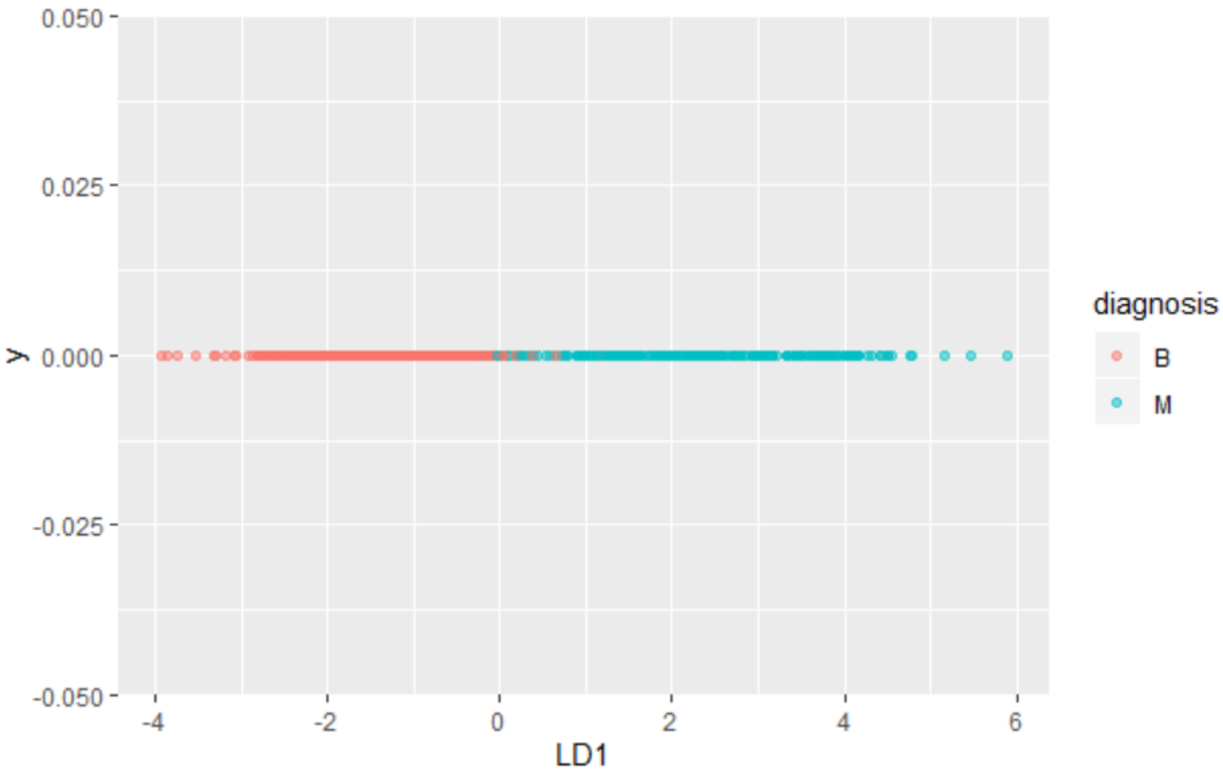


Figure 6 : Projection of WBCD features onto the axis with maximum class separability

After applying LDA, PCA we saved the results of both and then fitted different machine learning classification methods and then found the accuracy and compared them based on 3 different metrics. The metrics used were accuracy, sensitivity, and specificity.

Classification

Apart from SVM, Logistic Regression and MLP, we also tested out

1) Naïve Bayes

2) Random Forest

3) Adaboost

Section 5: Results and Discussion

Metrics used to evaluate models

We validated each model with 5-fold cross validation test. We also pick three metrics to evaluate our models - namely accuracy, sensitivity and specificity.

The reason behind using 3 metrics instead of just a single one is based on the type of error. The accuracy is a measure of the number of correct target prediction, but it doesn’t concern with incorrectly identified predictions. The motivation is that the model should predict a patient with cancer as a non-cancer patient (False Negative) and also it should not predict a patient without cancer as a cancer patient (False Positive). The first case can turn out to be fatal and the second case can lead to costly treatments for the subject. We selected the model according to the best accuracy along with the best specificity and sensitivity used to evaluate models

Confusion Matrix

True Positives	False Positives
False Negatives	True Negatives

Accuracy: $\frac{TP+TN}{TP+FN+TN+FP}$

Sensitivity/ TPR: $\frac{TP}{TP+FN}$

Specificity/TNR: $\frac{TN}{TN+FP}$

The evaluation metrics for each model is documented below

Model/Metric	Accuracy	Sensitivity	Specificity
Naive Bayes	92.94	90.48	94.39
LDA	94.12	85.71	99.07
LDA + Naive Bayes	95.29	88.89	99.07
ANN	92.94	88.89	95.33
Logistic Regression	95.88	92.06	98.13
PCA + Logistic Regression	95.29	92.06	97.20
LDA + Logistic Regression	95.29	88.89	99.07
LDA + ANN	95.88	90.48	99.07
Random Forest	95.88	92.06	0.9813
KNN + LDA	95.29	88.89	99.07
PCA + Adaboost	94,71	90.48	97.20
PCA + SVM	95.29	87.65	99.07
AdaBoost	95.88	93.65	97.20
PCA+SVM	95.29	88.89	99.07
KNN	95.88	88.89	100
KNN + PCA	95.88	88.99	100
PCA + ANN	96.47	93.65	98.13

LDA+SVM(L2)	96.47	91.06	98.06
SVM(L2)	96.47	92.06	99.07

Comparison of results:

- The paper used the batch normalization as the only way of preprocessing of the dataset. As explained by the feature density plot the individual features are not completely linearly separable on the other hand we used PCA and LDA which transforms the dataset into a dimensional space where they are linearly separable and hence the number of features reduced than the original
- We tried to fit a number of models than the original one in order to understand how the data reacts to different models with different probabilistic assumptions.
- The original paper only fitted models such as KNN, ANN, GRU+SVM, SVM, Linear regression, softmax classification. Out of all these models we used all the models except the GRU+SVM given it's worse performance out of all the models.
- As the reason of the failure of the GRU+SVM was the underlying linear separable distribution of the dataset we tried to fit SVM with not only linear but also polynomial kernel to handle some non-linearity with the data.
- As expected just like the paper we got very good accuracy with linear and polynomial SVM which had better training speed, accuracy and specificity compared to the other models.
- The accuracy of the results was less than the accuracy of the models given in the paper is attributed to the fact that the authors had more control over hyper-parameter tuning than

we did. This contributes to the lower absolute accuracy in our implementation but yet the results follow the same trend as given in the paper.

- We used not just accuracy but along with that specificity and sensitivity which are a measure of a model is performing in terms of True positive(TP) and True negative(TN) as they are equally important measures in the system.

Conclusion:

To conclude, we tried to come up with the classifier which can classify the cancer patients with high accuracy. To achieve this task we used the paper[1] as our reference for the development of the system. Along with what directed in the paper we followed the procedure of data analysis, feature engineering and model fitting and hyperparameter tuning for implementation of the system. The dimensionality reduction techniques such as PCA helped with the more linear separable bifurcation of data. Hence the models such as SVM and ANN performed well on the data. Hence we successfully recreated the paper results and did analysis to examine the accuracy shift for different probabilistic and nonprobabilistic methods.

References or Bibliography

[1] J. Ferlay, I. Soerjomataram, R. Dikshit et al., “Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012,” International Journal of Cancer, vol. 136, no. 5, pp. 359–389, 2014

[2] American Cancer Society. Breast Cancer Facts & Figures 2017-2018. Atlanta: American Cancer Society, Inc. 2017.

[3] Sharma GN, Dave R, Sanadya J, Sharma P, Sharma KK. Various types and management of breast cancer: an overview. J Adv Pharm Technol Res. 2010;1(2):109–126

[4] American Cancer Society, Cancer facts and figures 2019

[5] A. Cruz, D.S. Wishart, Applications of machine learning in cancer prediction and prognosis Cancer Informat, 2 (2006), p. 59

[6] Dana Bazazeh and Raed Shubair. “Comparative study of machine learning algorithms for breast cancer detection and diagnosis,” 2016 5th Int. Conf. on Electronic Devices, Systems and Applications (ICEDSA), 6-8 December 2016, Ras Al Khaimah, UAE.

[7] Chaurasia V., Pal., S, Tiwari., BB.: Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology, Vol. 12(2), pp. 119–126. DOI: http://dx.doi.org/10.1177/1748301818756225. (2018)

[8] Saxena., S, Burse., K.: A Survey on Neural Network Techniques for Classification of Breast Cancer Data. In; International Journal of Engineering and Advanced Technology, Volume-2, Issue-1, pp 234-237, (2012)