Building a predictive model for Breast Cancer Diagnosis using Machine Learning Techniques

Simranpreet Kaur
M.Sc Computer Science
University of Windsor
Student Id: 110005514
Email: kaur1a1@uwindsor.ca

Vipul Malhotra
M.Sc Computer Science
University of Windsor
Student Id: 105111504
Email: malho117@uwindsor.ca

***Abstract*—According to American Cancer Society(ACS), breast cancer (BC) is a group of diseases in which cells in breast tissue change and divide uncontrolled, typically resulting in a lump or mass. In 2019, an estimated 268,600 new cases of invasive breast cancer will be diagnosed among women and approximately 2,670 cases will be diagnosed in men. BC figures among the major causes of concern worldwide. According to the latest GLOBOCAN statistics [World Health Organization], it was the second most frequently diagnosed cancer and the fifth cause of cancer mortality worldwide, responsible for 6.4 percent of all deaths [3]. Classification and data mining methods are an effective way to classify data. Especially in medical field, these methods are widely used in diagnosis and analysis to make decisions. In this report, a performance comparison between different machine learning algorithms: Support Vector Machine (SVM), Decision Tree, Naive Bayes (NB) and k Nearest Neighbors (k-NN) [2] on the Breast Cancer Wisconsin [1] dataset is conducted. In short, the objective of this study is to build a predictive model that will improve the accuracy, objectivity and reproducibility of breast cancer diagnosis by Fine needle aspiration (FNA).**

***Index Terms*—Instance-based Learning, Non-parametric models, Supervised Learning, Principal Component Analysis (PCA), K-NN(Nearest Neighbor), Support Vector Classifier (SVC)**

I. INTRODUCTION

Breast cancer (BC) is one of the most common cancers among women worldwide representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today's society. Big data has advanced not only the size of data but also creating value from it. Big data - that becomes synonymous of data mining, business analytics and business intelligence, has made a big change in BI from reporting and decision to predicting results [4]. Data mining approaches for instance, applied to medical science topics rise rapidly due to their high performance in predicting outcomes, reducing costs of medicine, promoting patients health, improving healthcare value and quality and in making real time decision to save people's lives. There are many algorithms for classification and prediction of breast cancer outcomes. The present experiment gives a comparison between the performances of four classifiers: SVM, NB, Decision Tree and k-NN which are the most influential data mining algorithms in research community and are among the top 10 data mining algorithms. Our aim is to evaluate efficiency and effectiveness of these algorithms in terms of accuracy, sensitivity, specificity and precision.

II. PROBLEM STATEMENT

The problem in the dataset is the binary classification problem which involves a large number of features. The inputs are the characteristics of the given dataset (such as the radius, smoothness, compactness etc. of the distribution of cells), while the output is binary, i.e., benign or malign. So in order to understand the significance of the features, we have to perform some strategies for feature reduction, apply a binary classification algorithm and iterate this process, until performance saturates. To put it plainly, the objective of this study is to build a predictive model that will improve the accuracy, objectivity and reproducibility of breast cancer diagnosis by FNA.

III. RELATED WORK

Classification is one of the most important and essential task in machine learning and data mining. A lot of research has been conducted to apply data mining and machine learning on different medical datasets to classify Breast Cancer. Many of these experiments resulted in good classification accuracy.

Vikas Chaurasia and Saurabh Pal [7] compared the performance criterion of supervised learning classifiers; such as Naive Bayes, SVM-RBF kernel, RBF neural networks, Decision trees (J48) and simple CART to find the best classifier in breast cancer datasets. The experimental result showed that SVM-RBF kernel is more accurate than other classifiers; it scored accuracy of 96.84% in Wisconsin Breast Cancer dataset.

Djebbari et al. [13] considered the effect of ensemble of machine learning techniques to predict the survival time in breast cancer.Their technique showed better accuracy on their breast cancer data set when compared to previous results. S. Aruna and L.V Nandakishore [14], compared the performance of Decision Tree, Naive Bayes, Support Vector Machine (SVM) and K- Nearest Neighbor (K-NN) to find the best classifier in WBC. SVM proves to be the most accurate classifier with accuracy of 96.99%. Angeline Christobel. Y and Dr. Sivaprakasam [8], achieved accuracy of 69.23% using decision tree classifier (CART) in breast cancer datasets.

The accuracy of data mining algorithms SVM, IBK, BF Tree is compared by A. Pradesh [15]. The performance of MO has shown a higher value compared with other classifiers. T.Joachims [10] reaches accuracy of 95.06% with neuron fuzzy techniques when using Wisconsin Breast Cancer dataset.

With respect to all related work mentioned above, our work compares the behaviour of data mining algorithms SVM, NB, k-NN and C4.5 using Wisconsin Breast Cancer dataset in both diagnosis and analysis to make decisions. The goal is to achieve the best accuracy with the lowest error rate in analysing data. To do so, we compare efficiency and effectiveness of those approaches in terms of many criteria, including: accuracy, precision, sensitivity and specificity, correctly and incorrectly classified instances and time to build model, among others.

IV. METHODOLGY

A. Data Exploration & Exploratory Visualization

The dataset[1] has 569 rows and 33 columns. Amongst the 33 columns, the first two are ID number and Diagnosis (M = malignant, B = benign). And the last column is an unnamed column with only NaN values, so it is removed right away. The other 30 columns correspond to mean, standard deviation and the largest values (points on the tails) of the distributions of the following 10 features computed for the cell nuclei. All feature values are recorded with four significant digits. The class distribution of the samples is such that 357 are benign and 212 are malignant.

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
0	842302	M	17.99	10.38	122.80	1001.0	0.11840
1	842517	M	20.57	17.77	132.90	1326.0	0.08474
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960
3	84348301	M	11.42	20.38	77.58	386.1	0.14250
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030

Fig 1. Some features of the Wisconsin dataset

B. Data exploration

In data exploration, we will visualize the distribution statistics; apply appropriate data preprocessing steps, then look for the outliers and observe feature correlations that will guide in feature selection.

C. Data Preprocessing & Outlier Detection

In this step we are cleaning the dataset by dropping the outliers i.e. identify the outliers, remove them and observe the distribution again. It is to be noted that the features contain mean, se and worst values of the measurements of the 10 features. Since worst and se values determine the quality of measured data, we are observing only the last 20 features for removing outliers. In other words, we are trying to find in each of the features the observations that are lying away from their respective interquartile range. Each of these bad points are counted in a dictionary and finally bad points to be discarded are selected as those that occurred with highest frequency i.e., points that were bad in most features.

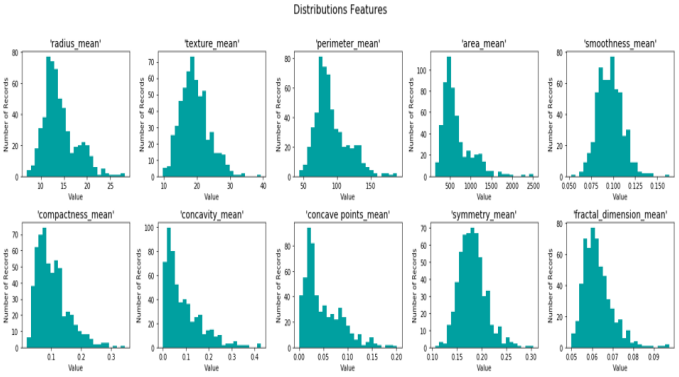

Fig 2. Dataset before removing the outliers.

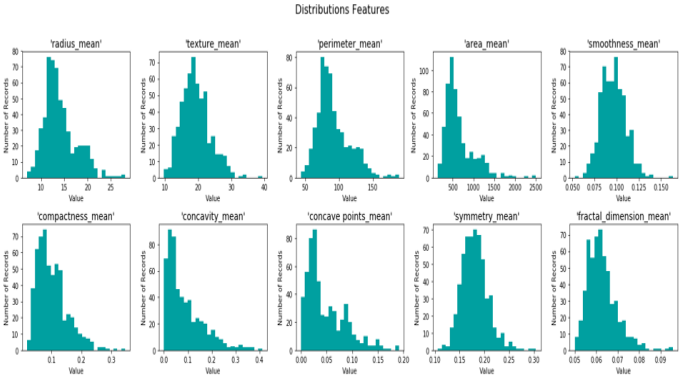

Fig 3. Dataset after removing the outliers.

From the Fig 2. And Fig 3., we can see that the distribution characteristics have changed especially w.r.t to the features depicting dimensions of the cell cultures. 2. We can now apply logarithmic transformations to features in order to remove the skewness and apply min-max scaling as well. Skewness has reduced in most plots except for some, then again look for outliers and clean the dataset by dropping the outliers if any and check for the distribution statistics. Below is a plot showing the size distribution of benign and malign samples, before and after data transformation operations.

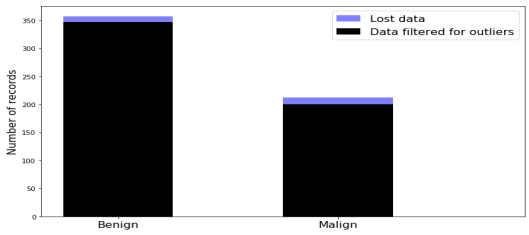

Fig 4. Bar plot of output classes.

D. Visualizing feature correlations

Let us now discover feature correlations if any. Almost all numbers in the correlation matrix are high indicating high correlation between the features. Radius, area and perimeter are the features that are correlated and this is established in the following correlation heat map.

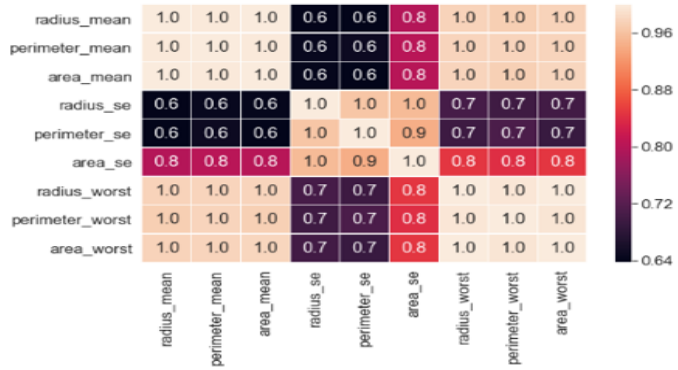

Fig 5. Correlation matrix for area features

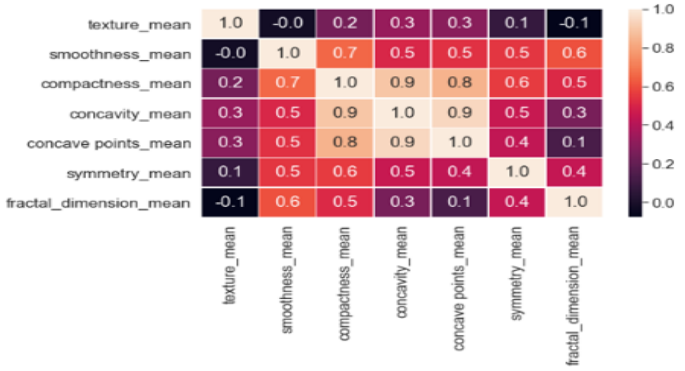

Fig 6. Corr. between compactness,concavity,concave_points in terms of mean

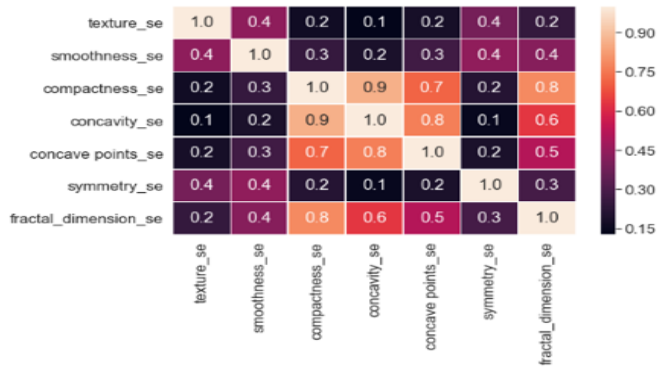

Fig 7. Corr. between compactness_se,concavity_se,concave_points_se

The final conclusions of feature selection implies that keeping one of the correlated features in each of the list item below might probably aid in classification such as:

- radius_mean, radius_se, radius_worst
- perimeter_mean, perimeter_se, perimeter_worst
- area_se, area_worst

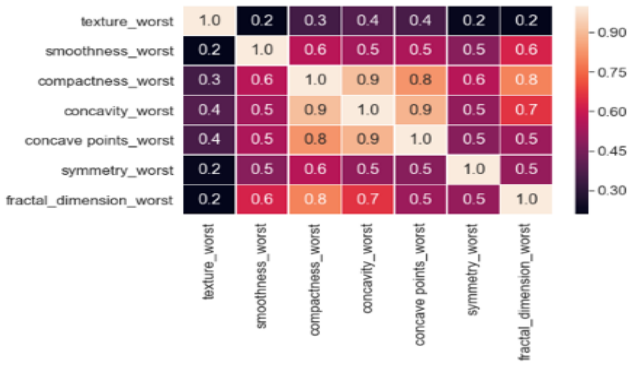

Fig 8. Corr. between compactness,concavity,concave_points in terms of worst

- smoothness_se, compactness_se, concave_points_se, concavity_se, symmetry_se
- compactness_worst,concavity_worst,concave_points_worst

So now the data has been scaled, transformed to reduce skewness, outliers have been removed, and feature correlations have been inspected. The next study would be feature transformation. By applying PCA, new dimensions that best maximizes the variance of features can be discovered. In addition to finding these dimensions, PCA also reports the captured variance of each dimension.

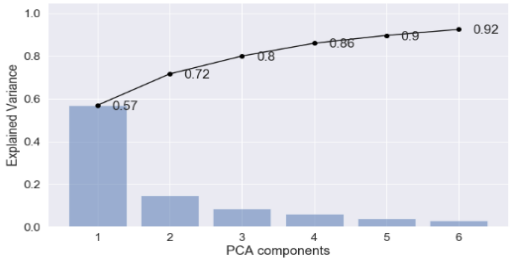

Fig 9. Graph between Explained VariancePCA Components

The Fig 9. indicates that 92% of variance in the data can be achieved with just 6 dimensions instead of the 30 features. PCA is important when dealing with data having a lot of features. It will minimize noise by only using the most important set of independent features.

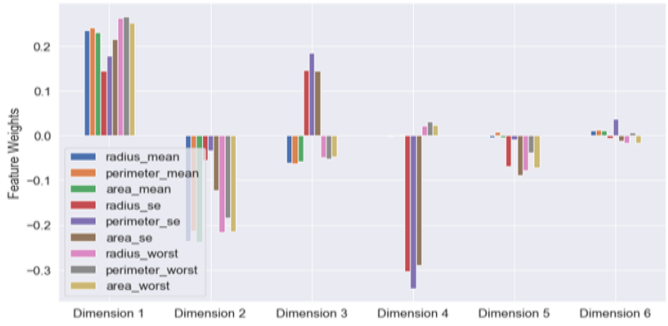

Fig 10. Plots for the feature weights and only dimensions

The above plots indicate the following:

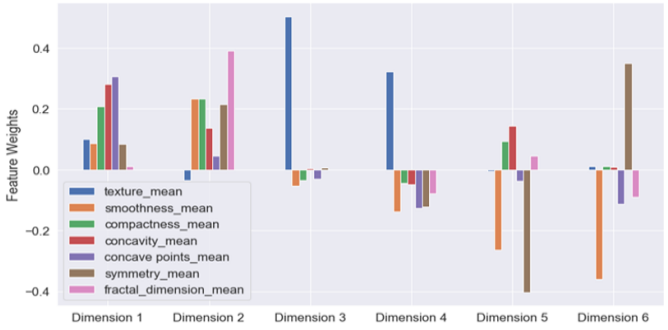

Fig 11. Plots for the feature weights and mean_non_dimensions

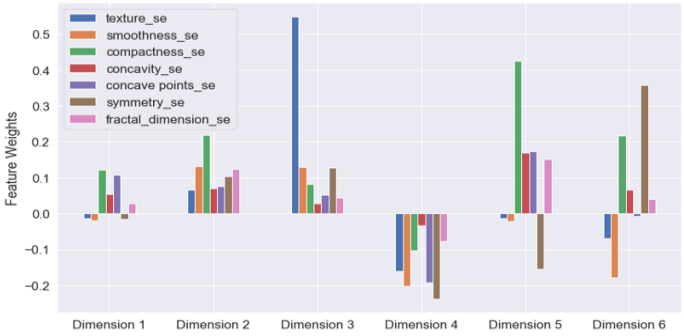

Fig 12. Plots for the feature weights and se_non_dimensions

1) The first principal component dimension has all positive weights.
2) The second principal component dimension has positive weights for all features except those related to cell dimensions (radius, perimeter, mean). The first two dimensions contribute to up to 72% variance.
3) texture_mean, texture_se and texture_worst all three contribute heavily to the third principal component dimension.
4) radius, perimeter, area ”mean” and ”worst” features contribute only to the first two principal component dimensions.

Below is a plot indicating that a good classification can be achieved with just 2 dimensions. It is not clear from the plots if adding the 3rd dimension improved classification. This can be known only by applying classification models to the reduced features and analyzing the scores.

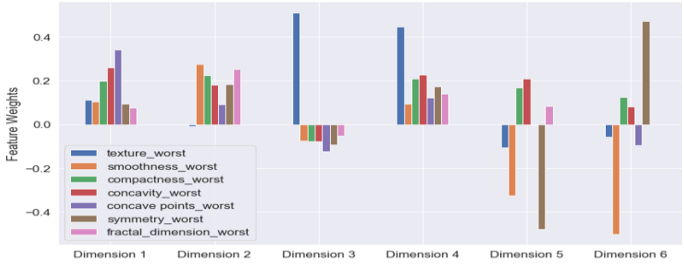

Fig 13. Plots for the feature weights and worst_non_dimensions

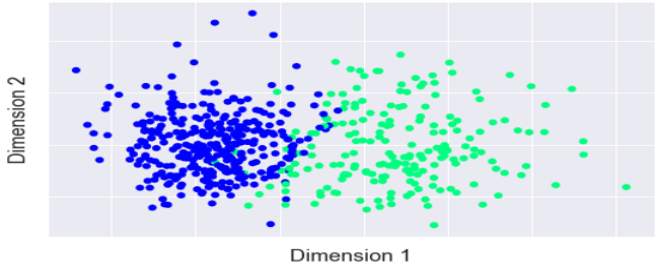

Fig 14. Plots for the feature weights and worst_non_dimensions

V. ALGORITHMS & TECHNIQUES

The reduced features from above will now be used in a binary classification algorithm. There are many algorithms suitable for the problem statement, but we conducted an experiment by choosing four classifiers: NB, Decision Tree, k-NN, SVM and have compared their performances. Let’s first briefly discuss the techniques being used in order to have an understanding of these algorithms.

1) **Gaussian Naive Bayes Classifier (GaussianNB)**

Naive Bayes method is a supervised learning algorithm based on Bayes theorem with the naive assumption of independence between various pairs of features. Gaussian Naive Bayes method assumes the likelihood of the features to be Gaussian. The advantage of this algorithm is that it is simple, fast and requires relatively small amount of training data. However the disadvantage with this algorithm is, it’s overly simplified assumption of independence between feature pairs. The problem in hand is a binary classification and the data we are using is the reduced feature matrix after applying PCA. This implies that the transformed features are independent of each other and thus it is likely that this algorithm might be successful in producing good results. However, if the feature-output is not so simple to be captured by this Naive algorithm, it definitely wouldn’t be suitable for our problem [6].

2) **Decision Trees/Random forests**

Decision tree is a non-parametric model that can classify data based on a tree of decision rules. Random forests on the other hand is an ensemble learning method, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. The advantage of this method is that they are simple to understand, visualize and interpret. It can handle both categorical and numerical variables without much pre-processing [12]. While the performance of a decision tree might improve with the number of rules set, it could very easily run into a problem of overfitting if the right questions are not asked. They are unstable to small variations in data and can often create biased trees if some classes dominate. Random forests on the other hand has the advantage of not falling into the issue of overfitting [11].

3) **K-Nearest Neighbors Classifier**

K-nearest neighbor's algorithm (k-NN) is a type of instance-based learning/lazy learning. It is also a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

- In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then voronoi diagram is constructed for its nearest neighbor and the object is simply assigned to the class of that single nearest neighbor.
- In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors.

The k-NN algorithm is the simplest of all machine learning algorithms. A peculiarity of the k-NN algorithm is that it is sensitive to the local structure of the data [7].

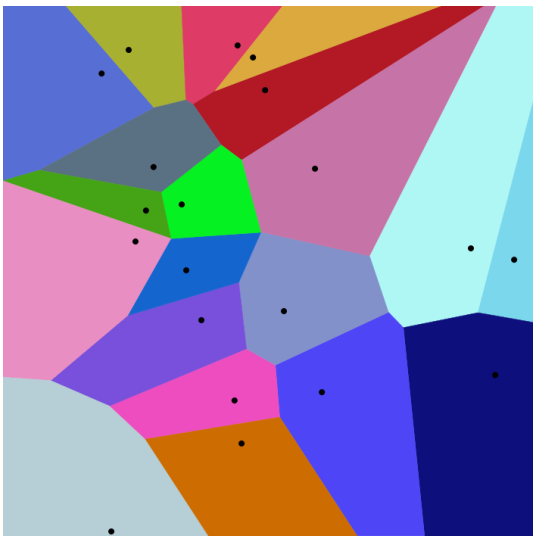

Fig 15. Representation of Voronoi Diagram in k-NN

4) **Support Vector Classifier**

Support vector machine constructs hyperplanes in infinite-dimensional spaces to achieve classification. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. The advantage of the method is very clear in the sense that it is a very formal approach to classification. The disadvantage, however, is the high training time complexity (more than quadratic with the number of samples) which makes it difficult to apply this algorithm to a couple of 10000 samples [5].

VI. EXPERIMENTS

Step 1: Begin with applying models (NB, Decision Tree, k-NN, and SVM) on the reduced data from above.

	GNB	RF	KNN	SVC
Average fit time	0.004s	0.018s	0.010s	0.003s
Average score time	0.005s	0.012s	0.025s	0.006s

Fig 16. Table displaying average fit and score time

The above plots leads to the following derivations:

1) Gaussian Naive Bayes is the least performing algorithm of all as suspected.
2) Random Forests have the best training scores (sometimes even perfect), but the testing scores are not so good. This indicates a biased model.
3) KNN and SVC are comparable in terms of training accuracy and fitting time as KNN training accuracy scores seem higher than SVC. But its precision scores are slightly lower than SVC. It can be observed that the variance of scores across the folds (as seen from the mean, min, max values) is lower in KNN for the test recall scores, but the behaviour is vice-versa for the test precision scores. KNN has low fitting time but higher score time and it is vice versa for SVC. However, in this case since the dataset is so small, it doesn't really matter [8]. We now print the f-beta test scores for KNN and SVC models to make the final decision on best model.

Step 2: The next step of the strategy is to remove some of the features observed from data exploration/visualization (i.e. feature selection), followed by PCA transformation and use of classifier models again on that to see if accuracy improves.

The results do not indicate any model improvement, rather there is a performance drop due to the selection of features before transformation. There is also an indication of overfitting, since training scores are much better, while testing score are worse.

Step 3: The final step for model refinement is fine tuning of the model hyper parameters.

The results indicate that all the algorithms give the same result. Only the k-neighbors parameter has any significant effect on the scores.

VII. RESULTS

Best obtained from grid search: Average fit time is: 0.004s and Average score time is: 0.028s

Second best model from grid search, it has lower variance of test scores across folds: Average fit time is: 0.003s and Average score time is: 0.022s

VIII. DISCUSSION/SUMMARY

We have addressed the binary classification problem of cancer diagnosis from FNA tests, based on the following strategy:

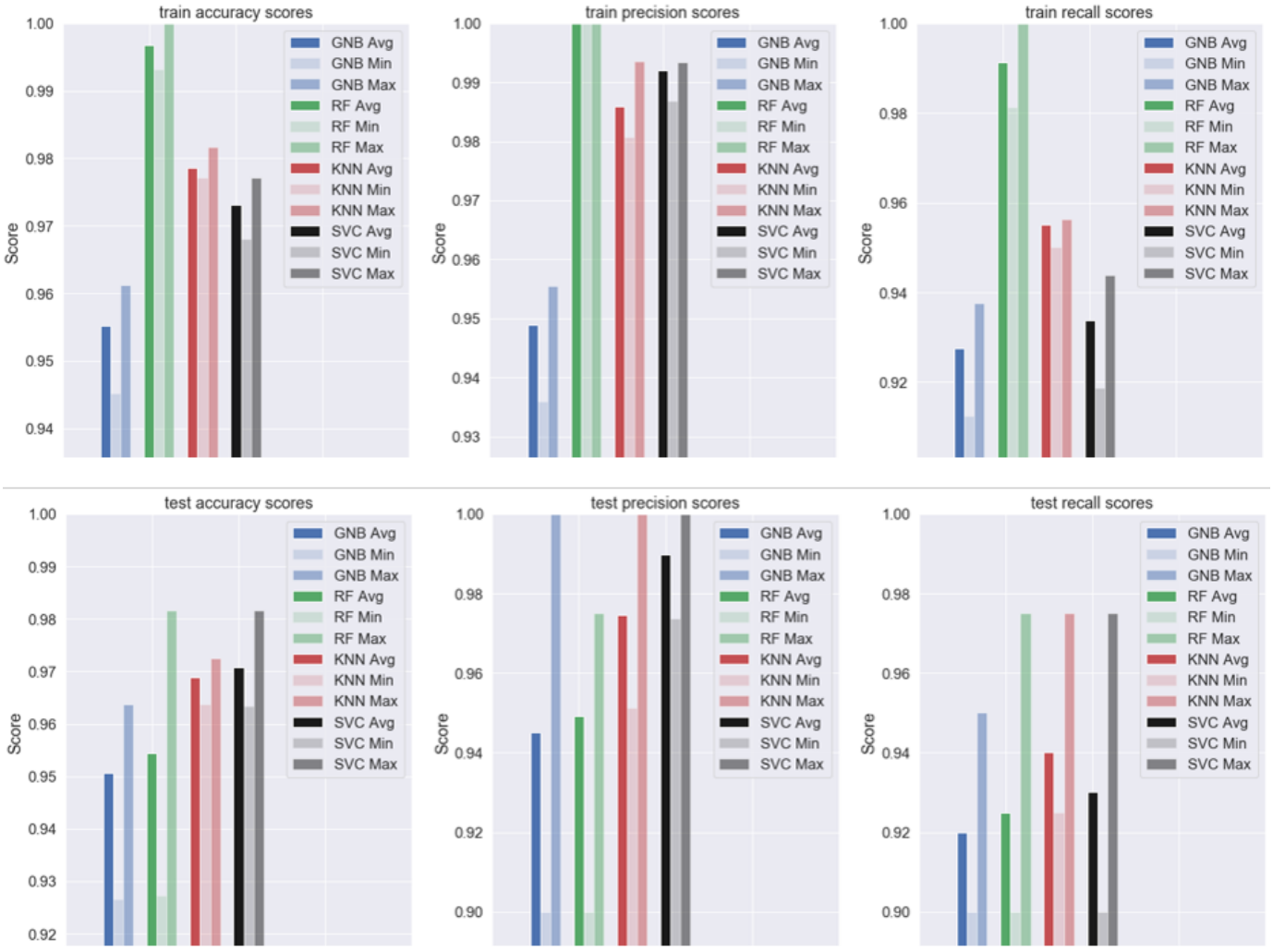

Fig 17. Plots displaying train and test accuracy, precision and recall scores

			Avg	Min	Max
KNN	train	f1	0.97	0.968	0.975
	test	f1	0.957	0.949	0.963
SVC	train	f1	0.962	0.955	0.968
	test	f1	0.959	0.947	0.975

Fig 18. Console displaying KNN and SVC train and test scores

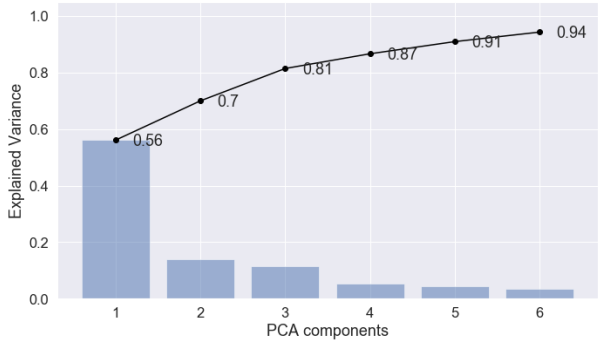

Fig 19. Graph between Explained Variance and PCA Components

1. We did an extensive data exploratory and visualization analysis of res constituting the test results.

- Log transformation and min max scaling was applied.
- Outliers were detected based on points lying outside the interquartile range. Points that were identified as outliers in the most number of features were dropped. This process resulted in losing about 4% of the data. And the loss was more or less equally distributed, thus maintaining the class balance.

2. Following the above analysis, we attempted feature transformation based on PCA. The feature weights composing the first 6 principal component dimensions were represented. And scatter plots of first two and three dimensions were also plotted to visualize the separation achieved with just 3 transformed features.

3. After feature transformation, we used the cleaned, reduced, transformed features in classification algorithms like Gaussian Bayes, Random Forests, (non-parametric) K-nearest neighbors and support vector classifiers. The classifiers scores were estimated across 5 cross validation folds, and the mean, worst

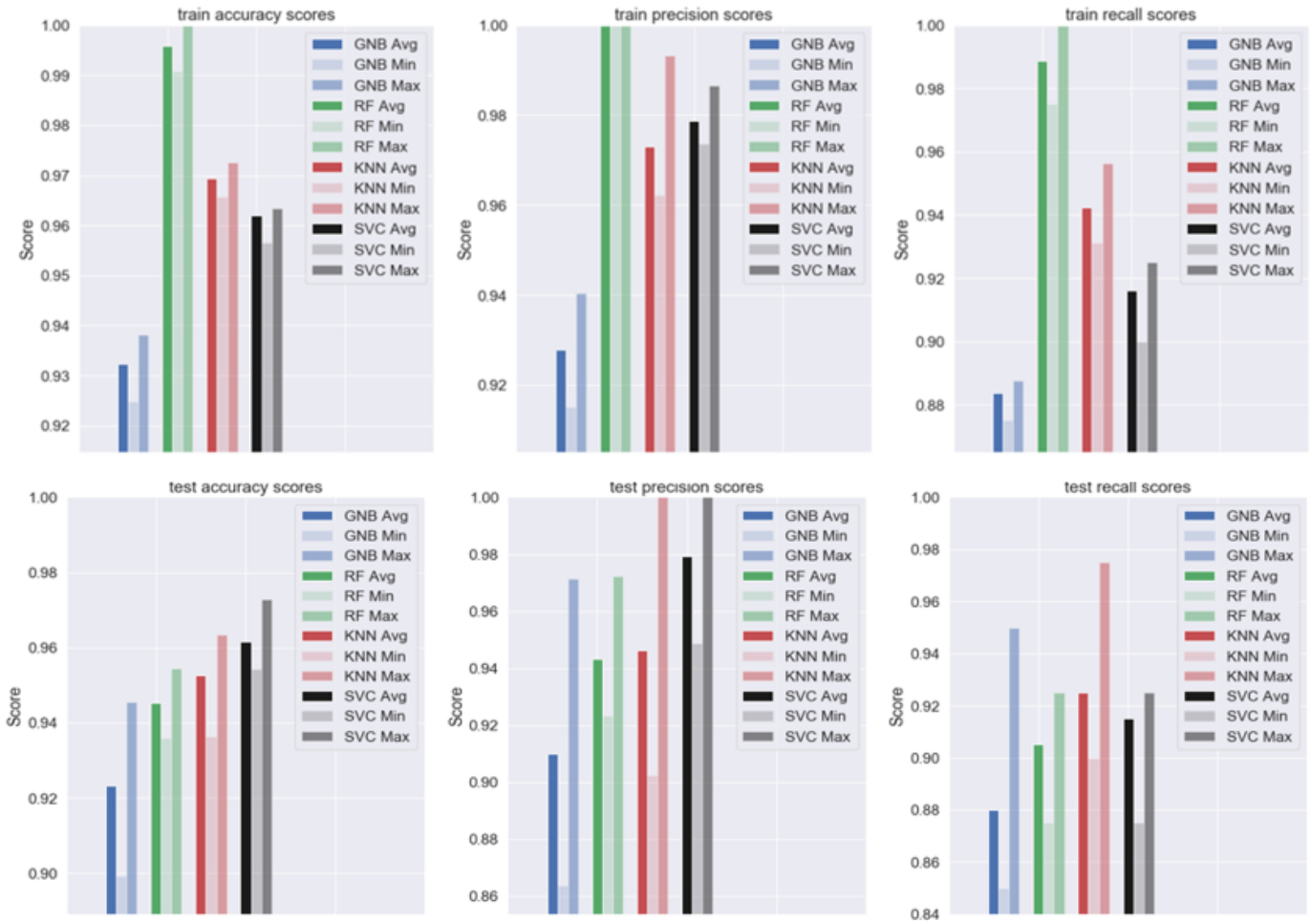

Fig 20. Plots displaying train and test accuracy, precision and recall scores

	mean_train_score	mean_test_score	param_algorithm	param_n_neighbors	rank_test_score	std_test_score
0	0.981039	0.980809	brute	4	1	0.012963
1	0.981039	0.980809	auto	4	1	0.012963
2	0.981039	0.980809	kd_tree	4	1	0.012963
3	0.981039	0.980809	ball_tree	4	1	0.012963
4	0.977460	0.969385	ball_tree	6	5	0.005567
5	0.977460	0.969385	kd_tree	6	5	0.005567
6	0.977460	0.969385	brute	6	5	0.005567
7	0.977460	0.969385	auto	6	5	0.005567

Fig 21. Table displaying different test scores

		Avg	Min	Max
train	accuracy	0.973	0.968	0.977
	precision	0.995	0.987	1
	recall	0.93	0.925	0.95
	f1	0.961	0.955	0.968
test	accuracy	0.973	0.955	0.991
	precision	0.995	0.973	1
	recall	0.93	0.9	0.975
	f1	0.961	0.935	0.987

Fig 22. Table for average fit and score time

		Avg	Min	Max
train	accuracy	0.979	0.977	0.982
	precision	0.986	0.981	0.994
	recall	0.955	0.95	0.956
	f1	0.97	0.968	0.975
test	accuracy	0.969	0.964	0.972
	precision	0.975	0.951	1
	recall	0.94	0.925	0.975
	f1	0.957	0.949	0.963

Fig 23. Table for average fit and score time

and best values of them were plotted to compare them. The KNN classifier came first among them on various aspects[9].

4. We also tried to drop features that were observed as not useful from the first step and then followed it with a PCA and classification. It seemed that dropping those features actually resulted in a small drop of the results.

5. Model hyper parameters were also optimized by a simple grid search algorithm.

6. The final test scores that we have are accuracy: 0.973, precision: 0.995 and recall: 0.93.

7. We have also printed out a second-best classifier which has slightly lower scores, but also lower variance of testing scores across the cross validation splits. However, in order to make

```
Confusion matrix of best model tested on original testing data:
    TP  FN
FN  70   0
TN   2  38


Confusion matrix of best model tested on testing data corrupted with false benign diagnoses, by 30%:
    TP  FN
FN  71   8
TN   1  30
```

Fig 24. Console displaying best obtained average fit and score times

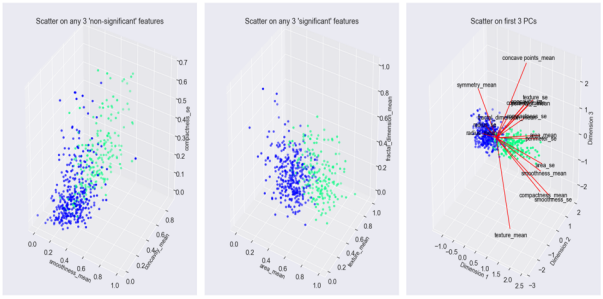

Fig 25. The three scatterplots displaying non-significant, significant and 3 PC's.

strong conclusions, it is definitely necessary to have more data points.

IX. CONCLUSION

To analyze medical data, various data mining and machine learning methods are available. An important challenge in data mining and machine learning areas is to build accurate and computationally efficient classifiers for Medical applications. In this study, we employed four main algorithms: SVM, NB, k-NN and DT on the Wisconsin Breast Cancer dataset. We tried to compare efficiency and effectiveness of these algorithms in terms of accuracy, precision, sensitivity and specificity to find the best classification accuracy. SVM reaches the accuracy of 97.13% and outperforms, all other algorithms. In conclusion, SVM has proven its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of precision and low error rate.

REFERENCES

[1] "UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set." Available:https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+

[2] Nikita Jain, Vishal Srivastava," Data Mining Techniques: a survey paper", IJRET: International Journal of Research in Engineering and Technology, 2013, 2321-7308, p. 116-119.

[3] 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.

[4] Asri H, Mousannif H, Al Moatassime H, Noel T. Big data in healthcare: Challenges and opportunities. 2015 Int Conf CloudTechnol Appl. 2015:1-7.

[5] Noble WS. What is a support vector machine? Nat Biotechnol. 2006;24(12):1565-1567. doi:10.1038/nbt1206-1565.

[6] Rish I. An empirical study of the naive Bayes classifier. IJCAI Work Empir methods Artif Intell. 2001;3(November):41-46.

[7] V. Chaurasia and S. Pal, "Data Mining Techniques : To Predict and Resolve Breast Cancer Survivability," vol. 3, no. 1, pp. 10–22, 2014.

[8] A. C. Y, "An Empirical Comparison of Data Mining Classification Methods," vol. 3, no. 2, pp. 24–28, 2011.

[9] A. Pradesh, "Analysis of Feature Selection with Classification : Breast Cancer Datasets," Indian J. Comput. Sci. Eng., vol. 2, no.5, pp. 756–763, 2011.

[10] Thorsten J. Transductive Inference for Text Classification Using Support Vector Machines. Icml. 1999;99:200-209.

[11] L. Ya-qin, W. Cheng, and Z. Lu, "Decision tree based predictive models for breast cancer survivability on imbalanced data," pp.1–4, 2009.

[12] Quinlan, J.R, "Induction of decision trees". Journal of Machine Learning ,1986, pp. 81-106.

[13] Djebbari, A., Liu, Z., Phan, S., AND Famili, F. International journal of computational biology and drug design (ijcbdd). 21st Annual Conference on Neural Information Processing Systems (2008).

[14] S. Aruna and L. V Nandakishore, "KNOWLEDGE BASED ANALYSIS OF VARIOUS STATISTICAL TOOLS IN DETECTING BREAST," pp. 37–45, 2011.

[15] A. Pradesh, "Analysis of Feature Selection with Classification : Breast Cancer Datasets," Indian J. Comput. Sci. Eng., vol. 2, no.5, pp. 756–763, 2011.