CMSC 691 - INTRODUCTION TO DATA SCIENCE

BREAST CANCER DETECTION (FEATURE SELECTION)

December 20, 2018

Ashish Deo
Prasad Akmar
University of Maryland, Baltimore County
Department of Computer Sciences

Breast Cancer Detection - Feature Selection

Ashish Deo, Prasad Akmar

***Abstract*—Cancer detection is one of the tasks machine learning can do very efficiently. In terms of machine learning, its a classification problem. Classification problems can be optimized for speed as well as accuracy using various techniques such as feature selection/elimination. We are observing effect of various feature selection techniques on three breast cancer dataset based on the accuracy of the model, and trying to identify significant features insights which may be useful for building combination of features across multiple datasets to solve this classification problem as well as help medical research on future data collection of the most significant features.**

I. Overview of Datasets

THIS section introduces the topic and leads the reader on to the main part.

Original	Diagnostic		Coimbra
• Observational Data by Dr. WIlliam H. Wolberg.	• Data from a digitized image of a fine needle aspirate (FNA)		• Data from a digitized image of a fine needle aspirate (FNA)
• Instances - 699 • Attributes – 10	• Instances: 569 • Attributes: 32		• Instances: 116 • Attributes: 10
• Clump Thickness • Uniformity of Cell Size • Uniformity of Cell Shape • Marginal Adhesion • Single Epithelial Cell Size • Bare Nuclei • Bland Chromatin • Normal Nucleoli • Mitoses	• Perimeter • Smoothness • Compactness • Concave points • Fractal dimension • Concavity • Symmetry	• Radius • Area • Texture	• Age • BMI • Glucose • Insulin • HOMA • Leptin • Adiponectin • Resistin • MCP.1

Fig. 1. Dataset description in tabular format.

II. Dataset Detail

A. Breast Cancer Wisconsin (Diagnostic) Data Set

1) Data Set Information: Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. A few of the images can be found at [Web Link]

Separating plane described above was obtained using Multi-surface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree Construction Via Linear Programming." Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, pp. 97-101, 1992], a classification method which uses linear programming to construct a decision tree. Relevant features were selected using an exhaustive search in the space of 1-4 features and 1-3 separating planes.

The actual linear program used to obtain the separating plane in the 3-dimensional space is that described in: [K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34]. This database is also available through the UW CS ftp server: ftp ftp.cs.wisc.edu cd math-prog/cpo-dataset/machine-learn/WDBC/

Data Set Characteristics:	Multivariate	**Number of Instances:**	569	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	32	**Date Donated**	1995-11-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	823935

Fig. 2. Wisconsin Diagnostic Dataset summary.

2) Attribute Information:

1) ID number
2) Diagnosis (M = malignant, B = benign)
3) 3-32) Ten real-valued features are computed for each cell nucleus:
   a) radius (mean of distances from center to points on the perimeter)
   b) texture (standard deviation of gray-scale values)
   c) perimeter
   d) area
   e) smoothness (local variation in radius lengths)
   f) compactness (perimeter^2 / area - 1.0)
   g) concavity (severity of concave portions of the contour)
   h) concave points (number of concave portions of the contour)
   i) symmetry
   j) fractal dimension ("coastline approximation" - 1)

3) Source:

1) Creators
   a) Dr. William H. Wolberg, General Surgery Dept. University of Wisconsin, Clinical Sciences Center Madison, WI 53792 wolberg '@' eagle.surgery.wisc.edu
   b) W. Nick Street, Computer Sciences Dept. University of Wisconsin, 1210 West Dayton St., Madison, WI 53706 street '@' cs.wisc.edu 608-262-6619
   c) Olvi L. Mangasarian, Computer Sciences Dept. University of Wisconsin, 1210 West Dayton St., Madison, WI 53706 olvi '@' cs.wisc.edu
2) Donor: Nick Street

B. Breast Cancer Wisconsin (Original) Data Set

1) Data Set Information: Samples arrive periodically as Dr. Wolberg reports his clinical cases. The database therefore reflects this chronological grouping of the data. This grouping information appears immediately below, having been removed from the data itself:

Group 1: 367 instances (January 1989)
Group 2: 70 instances (October 1989)
Group 3: 31 instances (February 1990)
Group 4: 17 instances (April 1990)
Group 5: 48 instances (August 1990)

Group 6: 49 instances (Updated January 1991)
Group 7: 31 instances (June 1991)
Group 8: 86 instances (November 1991)
Total: 699 points (as of the donated datbase on 15 July 1992)
Note that the results summarized above in Past Usage refer to a dataset of size 369, while Group 1 has only 367 instances. This is because it originally contained 369 instances; 2 were removed.
The following statements summarizes changes to the original Group 1's set of data:
Group 1 : 367 points: 200B 167M (January 1989)
Revised Jan 10, 1991
: Replaced zero bare nuclei in 1080185 & 1187805
Revised Nov 22,1991
: Removed 765878,4,5,9,7,10,10,10,3,8,1 no record
: Removed 484201,2,7,8,8,4,3,10,3,4,1 zero epithelial
: Changed 0 to 1 in field 6 of sample 1219406
: Changed 0 to 1 in field 8 of following sample:
: 1182404,2,3,1,1,1,2,0,1,1,1

Data Set Characteristics:	Multivariate	**Number of Instances:**	699	**Area:**	Life
Attribute Characteristics:	Integer	**Number of Attributes:**	10	**Date Donated**	1992-07-15
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	424875

Fig. 3. Wisconsin Original Dataset summary.

2) Attribute Information:

1) Sample code number: id number
2) Clump Thickness: 1 - 10
3) Uniformity of Cell Size: 1 - 10
4) Uniformity of Cell Shape: 1 - 10
5) Marginal Adhesion: 1 - 10
6) Single Epithelial Cell Size: 1 - 10
7) Bare Nuclei: 1 - 10
8) Bland Chromatin: 1 - 10
9) Normal Nucleoli: 1 - 10
10) Mitoses: 1 - 10
11) Class: (2 for benign, 4 for malignant)

3) Source:

1) Creators
   a) Dr. WIlliam H. Wolberg (physician) University of Wisconsin Hospitals Madison, Wisconsin, USA
2) Donor: Olvi Mangasarian (mangasarian '@' cs.wisc.edu) Received by David W. Aha (aha '@' cs.jhu.edu)

C. Breast Cancer Coimbra Data Set

1) Data Set Information: There are 10 predictors, all quantitative, and a binary dependent variable, indicating the presence or absence of breast cancer. The predictors are anthropometric data and parameters which can be gathered in routine blood analysis. Prediction models based on these predictors, if accurate, can potentially be used as a biomarker of breast cancer.

2) Attribute Information:

1) Quantitative Attributes:
   a) Age (years)
   b) BMI (kg/m2)
   c) Glucose (mg/dL)
   d) Insulin (U/mL)
   e) HOMA Leptin (ng/mL)
   f) Adiponectin (g/mL)
   g) Resistin (ng/mL)
   h) MCP-1(pg/dL)
2) Labels:
   a) 1 = Healthy controls
   b) 2 = Patients

3) Source: Miguel Patrcio(miguelpatricio '@' gmail.com), Jos Pereira (jafcpereira '@' gmail.com), Joana Crisstomo (joanacrisostomo '@' hotmail.com), Paulo Matafome (paulo-matafome '@' gmail.com), Raquel Seia (rmfseica '@' gmail.com), Francisco Caramelo (fcaramelo '@' fmed.uc.pt), all from the Faculty of Medicine of the University of Coimbra and also Manuel Gomes (manuelmgomes '@' gmail.com) from the University Hospital Centre of Coimbra

III. Exploring DataSet

A. Breast Cancer Wisconsin (Diagnostic) Dataset

- There are no Null values in the Dataset
- Datatype for all the columns in the dataset is int64
- Dataset has total 569 records with 357 Benign cases and 212 Malignant cases
- Dataset contains class label as B(indicating Benign case) and M(indicating Malignant case)
- Dataset was modified to add one more column(Label) to store the class label in a standard format as Benign or Malignant

Label	Count
Benign	357
Malignant	212

Fig. 4. Wisconsin Diagnostic Data distribution.

- Dataset comprises of 62.7% Benign cases and 37.3% Malignant cases, which is a good number for training and testing initial predictive model for Breast cancer
- Heatmap illustratring correlation among the column in the dataset can be plotted to identify and visible relation between the features in the dataset

B. Breast Cancer Wisconsin (Original) Dataset

- There are 16 Null values in the Dataset
- All the null values where present in the single column Bare Nuclei
- To handle the Null values in the dataset, these were replaced with mean(3.54) of all the values present in the column Bare Nuclei(mean value is : 3.544656).
- Datatype for all the columns in the dataset is int64

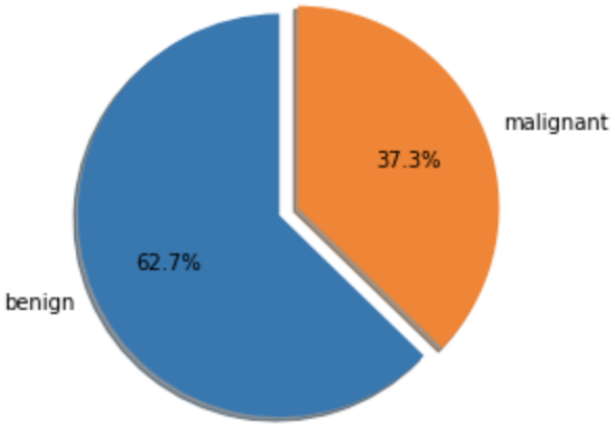

Fig. 5. Wisconsin Diagnostic Data distribution Pie chart.

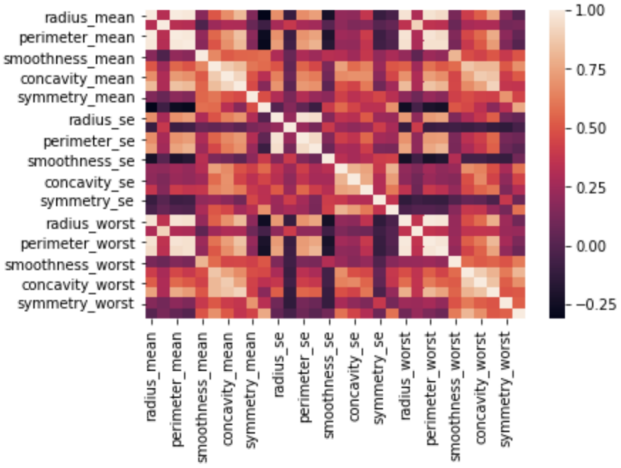

Fig. 6. Wisconsin Diagnostic Data distribution Heatmap.

- Dataset has total 699 records with 458 Benign cases and 241 Malignant cases
- Dataset contains class label as 2(indicating Benign case) and 4(indicating Malignant case)
- Dataset was modified to add one more column(Label) to store the class label in a standard format as Benign or Malignant

Label	Count
Benign	458
Malignant	241

Fig. 7. Wisconsin Original Data distribution.

- Dataset comprises of 65.5% Benign cases and 34.5% Malignant cases, which is a good number for training and testing initial predictive model for Breast cancer
- Heatmap illustratring correlation among the column in the dataset can be plotted to identify and visible relation between the features in the dataset

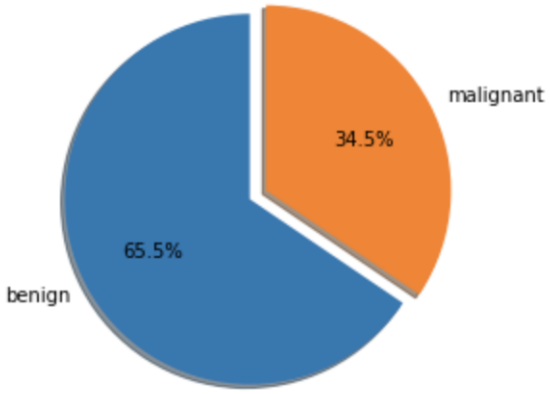

Fig. 8. Wisconsin Original Data distribution Pie chart.

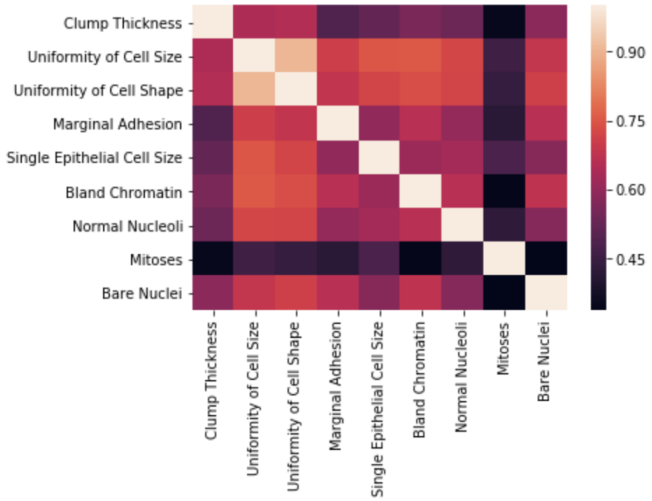

Fig. 9. Wisconsin Original Data distribution Heatmap.

C. *Breast Cancer Coimbra Dataset*

- There are no Null values in the Dataset
- Datatype for all the columns in the dataset is int64
- Dataset has total 116 records with 52 Benign cases and 64 Malignant cases
- Dataset contains class label as 1(indicating Benign case) and 2(indicating Malignant case)
- Dataset was modified to add one more column(Label) to store the class label in a standard format as Benign or Malignant

Label	Count
Benign	52
Malignant	64

Fig. 10. Coimbra Data distribution.

- Dataset comprises of 44.8% Benign cases and 55.2% Malignant cases, which is a good number for training

and testing initial predictive model for Breast cancer

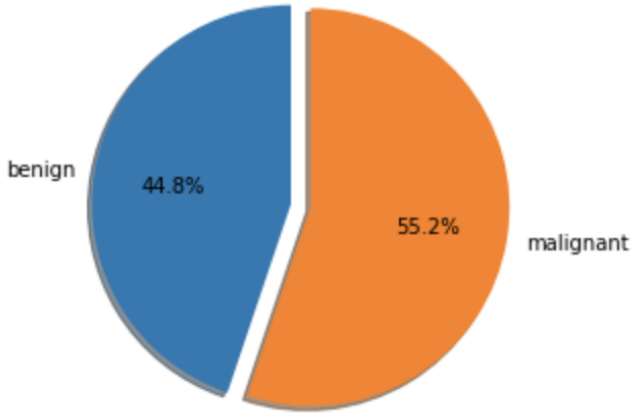

Fig. 11. Coimbra Data distribution Pie chart.

- Heatmap illustratring correlation among the column in the dataset can be plotted to identify and visible relation between the features in the dataset

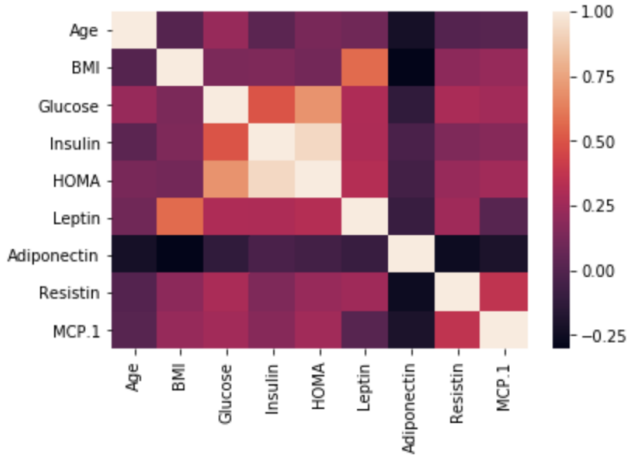

Fig. 12. Coimbra Data distribution Heatmap.

IV. Relation between dataset and why this project

- All the three dataset are addressing the similar classification problem of predicting a benign and malignant case of Breast cancer.
- Up until now, the available datasets on breast cancer data are very limited because limited number of instances recorded. We want to identify if the all the available datasets with same classification problem of positive and negative instances, can be combined in some way to get better accuracy at predicting.
- Feature selection in the classification problems can be really useful when it comes to datasets with high dimensionality of features. Feature selection can speedup the overall performance of the model and at the same time give better accuracy with less data.
- Dataset consists of different set of data features collected from different tests used for the diagnosis of Breast cancer.
- Dataset comprises of data for different set of patients, observed or extracted from totally different procedures or techniques for breast cancer detection.
- All the datasets individually aim to identify different Biomarkers for identification/detection of Breast cancer patient.
- Biomarkers are the significant features which helps to detect cases of Breast cancer patients from a group of individual, which may comprise of random combination of Benign and Malignant(ideally balanced cases of both benign and malignant) cases.
-
- The main relation between the three dataset can be derived from the fact that they intend to address classification problem of Breast cancer prediction.
- Since they comprise of data from different procedures or observational activities, there can not be any direct linkage identified between these dataset.
- But having said that, since they address the same common problem and comprises of data from different diagnostic tests for breast cancer detection, we say that if we can identify significant set of features from these dataset and combine them into a single feature set, we are expecting to predict benign and malignant cases of Breast cancer more precise and accurately.
- Early detection is the key for Treatment of Breast Cancer.
- To address the challenge of early detection of Breast cancer, data from one specific type of test is not adequate or sufficient to arrive to a conclusion with an accurate prediction.
- Features from different datasets, obtained through various diagnosis, screening and monitoring tests, can be used to identify the most significant biomarkers across the different dataset and then group them into a single set of features which can provide accurate and precise prediction of Breast cancer cases.

V. Activities and Results

- Initially we were trying to establish a direct relation between the datasets
- Later on analysing the dataset and identifying the source of the dataset, we came to conclusion that it was difficult to establish a direct relation between the datasets as they comprised of data from different screening, monitoring and diagnostic tests.
- Some of the data in the dataset was merely observational data, which was directly provided by the medical practitioner attending the patients.
- Such data cannot be considered reliable and cannot be considered for establishing the direct relation between dataset.
- Other data was obtained from totally different procedures and hence it is difficult to obtain same/similar data

from different type of procedure. For instance, wisconsin diagnostic dataset is obtained from images of a biopsy procedure on the other hand, original dataset comprises of mere observational data provided by Dr. WIlliam H. Wolberg for his patients.

- These two dataset cannot comprise similar significant data/feature that can be used to establish a direct relationship between different datasets.
- Second challenge was to identify if taking intersection of the derived significant feature set from different dataset was a better approach or taking union.
- Taking Union of derived significant features would result in increase in the final feature set, which we are calling as most significant features for Breast cancer prediction.
- Since union operation may result in addition of extra features in the final most significant feature set, we observed there was a decrease in the prediction accuracy when these feature sets were individually used for prediction.
- Based on the observation of initial decrease in the prediction accuracy with increased number of features due to union operation, we decided we will form the final most significant feature set by taking intersection of the intermediate feature set from different feature selection algorithm.
- Despite the first instinctive believe, we found out that model started to produce better result with the new feature set from intersection operation.
- This is when we changed the whole procedure of final feature selection from union operation to intersection operation.
- Following are the observed results from the all the three feature selection on three datasets.

1) Wisconsin Breast Cancer Dataset (Diagnostic)

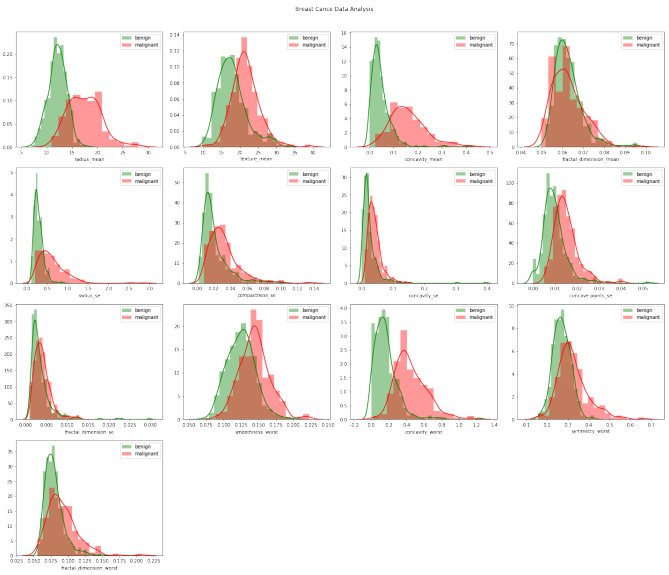

Fig. 13. Wisconsin Diagnostic Data Correlation plot.

2) Wisconsin Breast Cancer Dataset (Original)
3) Coimbra Breast Cancer Dataset

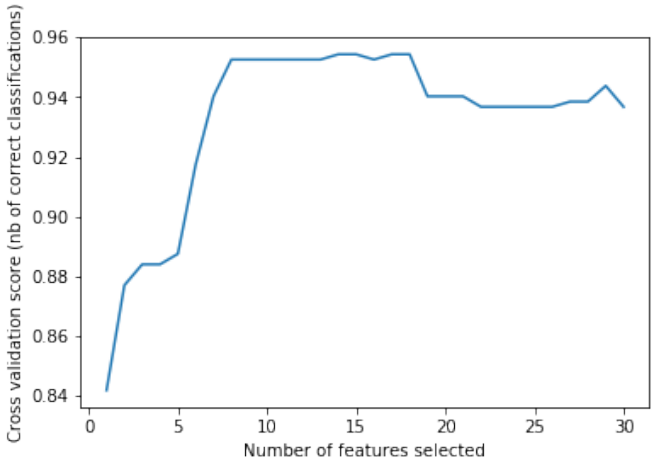

Fig. 14. Wisconsin Diagnostic Data RFECV.

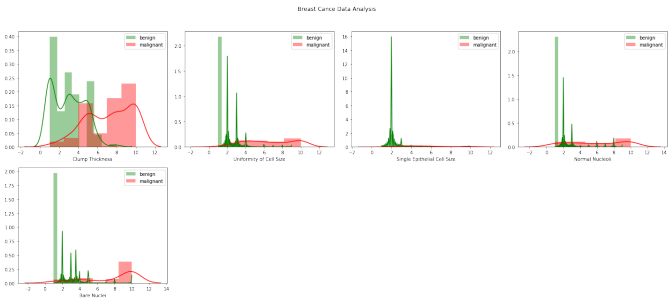

Fig. 15. Wisconsin Original Data Correlation plot.

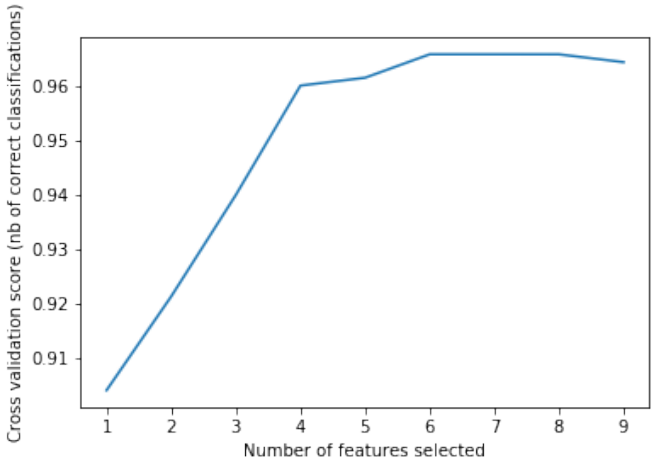

Fig. 16. Wisconsin Original Data RFECV.

VI. Process Flow

- We first started with Data Cleansing and Data Integration process to remove all unwanted and unnecessary features from the dataset like id, sequence number.
- Then we ran SVC classifier on these dataset to get an idea of how the model performs with the given dataset for classification task. The reason we chose it because we want to get the accurate classification to the problem. And SVM performs better compared to other classification

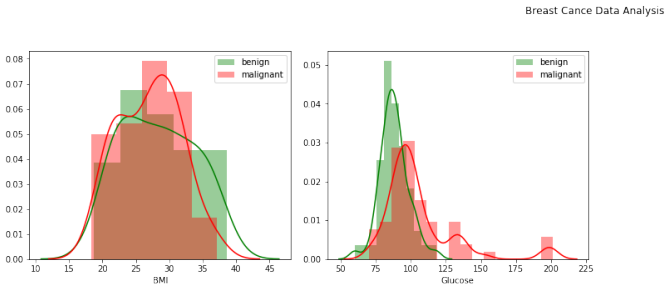

Fig. 17. Coimbra Data Correlation plot.

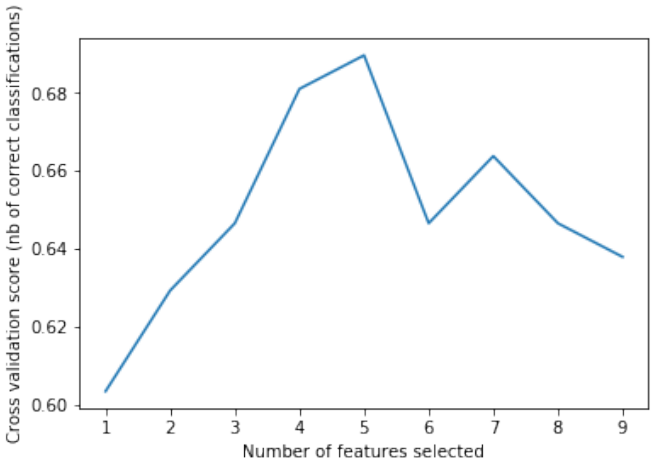

Fig. 18. Coimbra Data RFECV.

algorithms because of its ability to form unnatural shaped clusters.

- We recorded the accuracies(as tabulated below) for the dataset with SVC model using Gaussian kernel.

Dataset	Accuracy
Wisconsin (Diagnostic)	71.05
Wisconsin (Original)	92.14
Coimbra	70.8

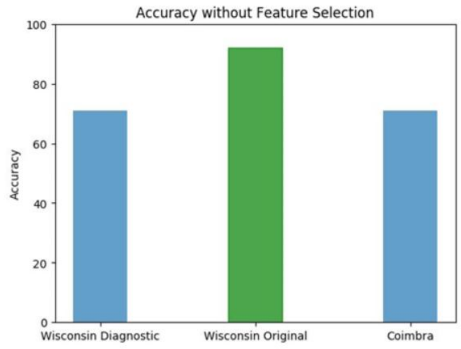

Fig. 19. Accuracy without Feature selection.

- Then we considered one dataset at a time and preformed feature selection operation on that dataset to obtain significant feature list for that dataset.
- Here we are using three different feature selection algorithms for research purpose to analyze the effect and performance of predictive algorithm on features selected from different feature selection algorithm. We have used Correlation with p-value, Recursive Feature Elimination and Select False Discovery Rate feature selection algorithms

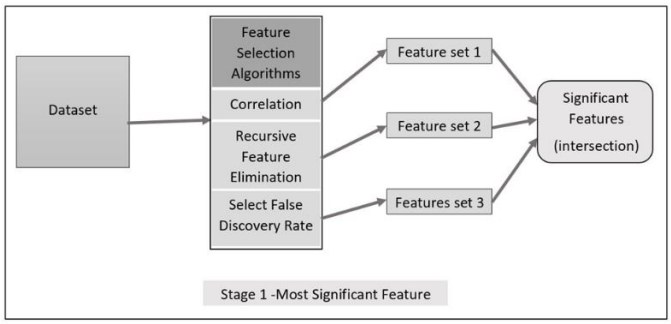

Fig. 20. Process Flow.

- Correlation with p-value:
  - Correlation is one of the most effective and popular method used for feature selection. Correlation uses feature elimination. Lets see how it works.
  - Correlation, as name suggests, is just correlation between two different features of the dataset.
  - For feature selection it can be used effectively because highly correlated features in a dataset will have same impact on the result if there is increase or decrease in the value. So one of the highly correlated features are removed from the dataset, which does not impact the accuracy of the model but increase the performance for sure.
  - P-value is is probability value for given statistical model if the null hypothesis is true, a set of statistical observations more commonly known as the statistical summary is greater than or equal in magnitude to the observed results. Removal of different features from the dataset will have different effects on the p-value for the dataset. We can remove different features and measure the p-value in each case. These measured p-values can be used to decide whether to keep a feature or not.
- Recursive Feature Elimination with Cross Validation:
  - Recursive feature elimination, as name suggests eliminates the features which least affect the result of the classification.
  - A feature is removed in every cycle and the accuracy is calculated.
  - This process is performed recursively and monitored. If the accuracy of the model reduces compared to the previous iteration, the feature is added again to the model.
- Select Fdr:
  - The false discovery rate (FDR) is a method of conceptualizing the rate of type I errors in null hypothesis testing when conducting multiple comparisons. FDR-controlling procedures are designed

to control the expected proportion of "discoveries" (rejected null hypotheses) that are false (incorrect rejections).
  - Select Fdr is based on False Discovery Rate.
- We will then obtain 3 different feature subset for a dataset, generated by executing different feature selection algorithms on the dataset
- We will then take intersection of all the selected feature set to generate most significant feature set for one dataset
- We consecutively performed feature selection and intersection operations on the dataset and ran a SVC with Gaussian kernel classifier to compute accuracy of model with existing/modified feature set.
- Accuracies for all the intermediate stage is as shown below in a tabularised format for better visualization.

VII. Insights and Knowledge

Following are some useful insights we gained from our experiments on the data.

- The charts and the bar graphs represents the effect of feature selection on all the three datasets.
- We considered Breast Cancer (Diagnostic), Breast cancer (Original) and Coimbra university hospital datasets and applied three different feature selection methods on the data. We ran SVM on all the resultant datasets and observed below results.

	No Feature selection	Correlation with p-value	RFECV	SelectFdr	Intersection
Wisconsin (Diagnostic)	63.15	93.85	92.1	61.4	87.71
Wisconsin (Original)	95	95.71	95.71	95.71	95
Coimbra	58.33	70.83	75	66.66	'0.83 (eliminating SelectFdr)

Fig. 21. Comparision Summary.

- As we can see, for the Wisconsin Breast Cancer Dataset (Diagnostic), the number of features the accuracy of SVM increases significantly with feature selection using using correlation and RFECV methods. When we performed the intersection on the resultant featuresets from three methods, the accuracy increased significantly as compared to the dataset without any feature selection. We observed that the number of features initially for this dataset were 30, and each method chose close to 18 features resulting in increase in the accuracy over all in the intersection of the three result sets.
- The intersection of the the three methods only resulted in 3 features. And even with the three features, the accuracy was significantly increased.
- For Wisconsin Breast Cancer Dataset (Original), we observed no significant gain or loss in the accuracy of SVM with of without performing feature selection. The number of features in the dataset initially were 10, and each feature selection method chose close to 8 unique features. The intersection of the the three methods only resulted in 3 features. And even with the three features, the accuracy was almost constant.
- The Breast Cancer Coimbra Dataset showed rather some interesting insights and different from the other two datasets. The number of instances in this dataset were very low, 116. So as we split the data 80:20 for training: testing, the number instances on which the data was trained fell down even further. But the overall accuracy of the model on this dataset after feature selection increased. Although, only one feature remained after we tried to take the intersection of the three feature selection methods. As the model trained on just one feature would not be any accurate, we chose to eliminate correlation feature set as it had only two features.
- Out of all the datasets, the maximum accuracy gains were observed in Breast Cancer Coimbra Dataset overall. And maximum accuracy gain for one individual method method were observed in Wisconsin Breast Cancer Dataset (Diagnostic).

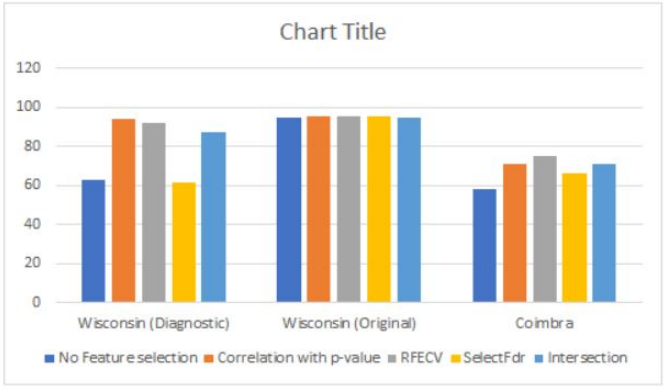

Fig. 22. Comparision Summary Bar Graph for Dataset.

- The above bar chart shows insightful information about the accuracies measured for all three datasets over all 4 methods, viz. No feature selection, Correlation with p-value, RFE-CV and SelectFdr. We cannot really infer any particular pattern from this as all the methods and intersection seem to affect differently to each dataset.

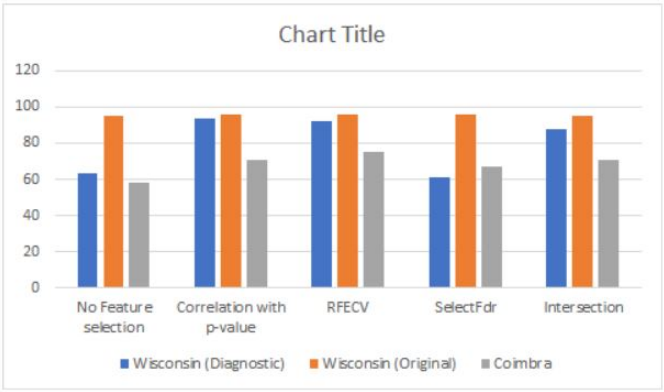

Fig. 23. Comparision Summary Bar Graph for Steps.

- The above bar chart combines the accuracies of SVM over the feature elimination method applied across all the datasets. Out of all, we could conclude from the above results that RFE-CV performs the best with consistent results closely followed by correlation and the intersection.
- Overall, the intersection of the different methods seems to perform well above average as per our expectation. Though it may not help much in increasing accuracy for the models which are already performing better. But it significantly increased the performance of the models with subpar accuracies otherwise.
  - Based on the activities and task done in this project, it is safe to say that different feature selection model gives different result.
  - For example, Correlation method of feature elimination is useful when there is good number of features available in the dataset.
  - Since correlation is a feature elimination technique, it needs the initial dataset to have good number of features so that after elimination, there is still adequate amount of features remaining to train and test the prediction model.
  - If we use correlation feature elimination technique on a dataset with less number of feature(less than 15 or 10), then the resulting features selected will only have fewer features in the set to run the model on. And due to the fewer number of features, model may not perform as expected in certain scenarios.
  - We were not able to use RFE without cross validation, as it was nearly not feasible to identify the best number of features to keep in the significant feature set. Using RFECV provided the flexibility of not specifying the number of significant features in advance and also provided with a strategical and explanatory trend for justifying the selection of appropriate number of features from the features set.
  - In most of the RFECV performed better than selectFDR and correlation feature elimination techniques to derive the final most significant feature set

VIII. Future Scope

- Due to unavailability of good amount of data, considering the fact that it is collected for Breast cancer diagnosed and detected patients, it was not feasible to collect data for the finally derived most significant feature set.
- Based on the observations and finding submitted in this project, if we can identify some benign and malignant patients for Breast cancer and perform data collection for the finally derived most significant features, we can use the data to train some classification model to predict Breast cancer in patient.
- Based on the accuracies of different classifier models, we can observe and validate the claim that using most significant features for Breast cancer prediction obtained across different biomarkers collected from different diagnosis, screening and monitoring test, will produce more accurate and precise predictions.

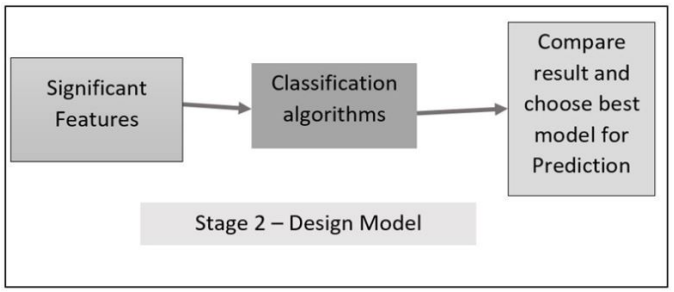

Fig. 24. Future Scope Process Flow.

- Here we have used SVM Classifier through out this project but we can also try this with some other classifiers to identify different patterns and accuracy in prediction.
- Random Forest is intrinsically suited for multiclass problems, while SVM is intrinsically two-class. For multiclass problem you will need to reduce it into multiple binary classification problems.
- Random Forest works well with a mixture of numerical and categorical features. When features are on the various scales, it is also fine. Roughly speaking, with Random Forest you can use data as they are. SVM maximizes the ”margin” and thus relies on the concept of ”distance” between different points. It is up to you to decide if ”distance” is meaningful. As a consequence, one-hot encoding for categorical features is a must-do. Further, min-max or other scaling is highly recommended at preprocessing step.
- If you have data with n points and m features, an intermediate step in SVM is constructing an nn matrix (think about memory requirements for storage) by calculating n2 dot products (computational complexity). Therefore, as a rule of thumb, SVM is hardly scalable beyond 10$\hat{5}$ points. Large number of features (homogeneous features with meaningful distance, pixel of image would be a perfect example) is generally not a problem.
- For a classification problem Random Forest gives you probability of belonging to class. SVM gives you distance to the boundary, you still need to convert it to probability somehow if you need probability.
- For those problems, where SVM applies, it generally performs better than Random Forest.
- SVM gives you ”support vectors”, that is points in each class closest to the boundary between classes. They may be of interest by themselves for interpretation.

References

[1] http://www.aaai.org/Papers/ICML/2003/ICML03-111.pdf
[2] https://www.lri.fr/~pierres/donn\%E9es/save/these/articles/lpr-queue/hall99correlationbased.pdf
[3] https://www.sciencedirect.com/science/article/pii/S1053811911013486
[4] https://www.sciencedirect.com/science/article/pii/S0045790613003066
[5] https://seaborn.pydata.org/examples/many_pairwise_correlations.html
[6] https://pandas.pydata.org/pandas-docs/stable/visualization.html
[7] https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.plot.html