Breast Cancer Diagnostic: Benign or Malignant

Alhussain Almarhabi

Abstract-Focused on designing classifier for breast cancer tumor if it is benign or Malignant using MCI Machine leaning dataset, established by university of Wisconsin. After applying expletory data analysis on the datasets, two methods – k nearest neighbors (KNN) and support vector machines (SVMs) - were applied to achieve the classification goals. Further investigation were done on the two methods to improve the classifier. The main concern was to improve the performance as much as possible and to reduce error type II (FN) which classify the patients as been healthy (Benign) while cancer classification is positive. We achieved 97% accuracy using SVMs and 96% accuracy using KNN.

I. INTRODUCTION

Cancer is considered to be one of top deadly diseases; Breast cancer is one type of cancer that start spreading in the breast or chest area, it's common in women and also affect men as well. The project goal is to classify whether the breast cancer, more specifically the tumor in breast area, is either benign for not cancerous or malignant for cancerous. After studying the database and analyzing its data, we can design a classifier to classify whether the cancer is benign or malignant given certain features. We have up to 32 features that we can use and investigate which one will contribute to the classification. The dataset features were extracted from a digitized image from a fine needle aspirate (FNA) of a breast mass as described by UCI Machine Learning Repository. The dataset has 569 instances 'patients found to have tumor in breast area' with the total of 32 attributes. This dataset was made by collaboration between clinical sciences center and computer sciences department at university of Wisconsin. The dataset has 10 main extracted attribute grouped in three parts mean, standard error and worst (largest). So the total features we have are 30 features and one column indicate the cancer type.

II. LITERATURE REVIEW

Breast cancer

First let us cover some important concepts a bout breast cancer. breast cancer is very common in women but it also can be present in men. about 1 in 8 U.S. Women (about12.4% 12 out of every 100) will develop invasive breast cancer and for men in lifetime risk of breast cancer is about 1 in 1,000 [1]. Cancer is developed when cells grown out of control and crowd out normal cells [2]. This makes it hard for body to work the way it should. Cancer nowadays can be treated very well if diagnosed early. There are many types of cancer, categorized base on spreading location 'metastasis', one of them is the breast cancer. Cancers are similar in the way they start spreading but they are different in how fast or slow they grow and spread and in addition how they respond to treatment whither it was surgery or chemotherapy [3]. Most cancer form a lump know as a tumor but not all tumor are cancer. Doctors take biopsy 'sample' of the lump and test if it's benign (not cancer and not dangerous) or malignant (cancer). Doctors use fine needle aspiration (FNA) , a very thin hollow needle attached to syringe to withdraw a small amount of tissue from a suspicious area [4].Doctors usually do not know the reasons why patient get cancer.[5]. Breast cancer refers to a malignant tumor that has developed from cells in the breast and it caused by genetic abnormality [6]. There are about nine stages of breast cancer where doctors can identifying the level of spread of cancer in breast.[7] 80 to 85 percent of breast lumps or tumors are benign especially in women younger than age 40 [6].

Related approach

There is well written decent research paper in 2017 that work on classification of breast cancer using similar dataset and different. One paper focused on using neural network architecture for breast cancer detection and classification. Papered proposed low-complexity back propagation and used a feed-forward networks for the fact of its simplicity at implementation in CMOS circuit with the same parameters as the original one [8]. The objective was to find the optimal activation function that minimizes the error of the classifier which found to be logsigmoid or hyperbolic tangent without biases. Another paper Wisconsin dataset of breast cancer, the paper investigate many classification techniques

and compared them in term of accuracy e.g. Naive Bayes, SVM, KNN, Random Forest and three more. The focus on the mean of the features and tested the accuracy for each algorithm with different test and train data percentage. The highest accuracy were 98% for SVM at 85% of data as training for SVM and 98.06% for CNN[9] . A group of researcher used Bayesian linear discrimination analysis (BLDA) for the breast cancer and achieved 83.45% accuracy [10]. Also, one approach used biomarkers to detect features known as micro ribonucleic acid (MicroRNA) with Naive Bayes classifier and Multi-layer perceptron that achieves more than 90% using all feature set [11].

III. MOTIVATIONS

There are three motivations to choose and work on this problem. First, this kind of project help us understand the humans body more by understanding small part of how current feature reflect on different types of tumor. Second, contribute to treat other related health problem. As soon we be able to establish good understanding of current issues we can use it to other cancer problem and even different diseases. Lastly, Curiosity! To see if we can apply our knowledge to application related to humans body which is difficult already.

IV. EXPIATORY DATA ANALYSIS - EDA

Before choose and apply our methods, it is better to perform expiatory analysis on the data to understand the dataset, look for pattern, clean any missing information and see it reflect the real life statistics. First, we did count plot for the two tumor classes which found that the dataset have 350 tumor classified as benign and 219 as malignant. Figure 1 shows the count plot of diagnosed tumor; this figure reflect the real world statistics which implies benign tumors are more common diagnosed.

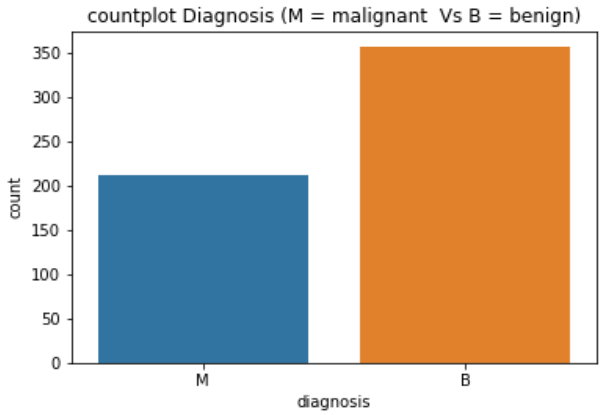

Figure 1: Countplot of tumors diagnosis classes.

Second, we did plot some features probability density function using kernel density estimation. We plot the pdf for radius_mean, radius_se and compactness_mean. Figure 2 shows the pdf of the three selected features and in it we can see how most of data overlap between the two classes. A better to show that clearly is by plotting pair plot between features for each group of the mean, standard error and worst, Figure 3 shows pair plot of worst group.

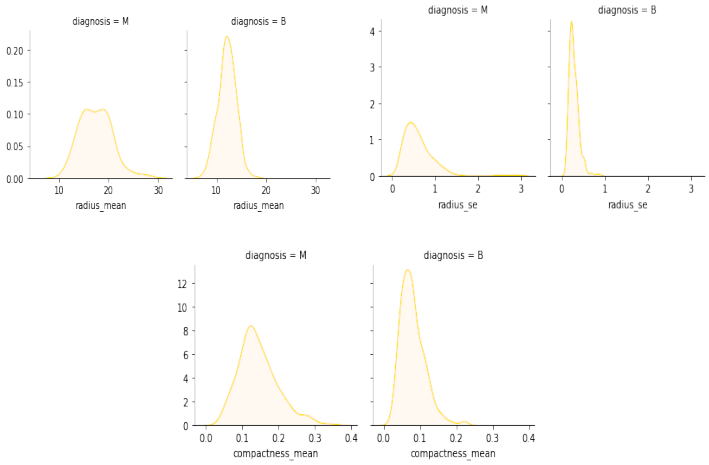

Figure 2: Kernel density estimation of selected features

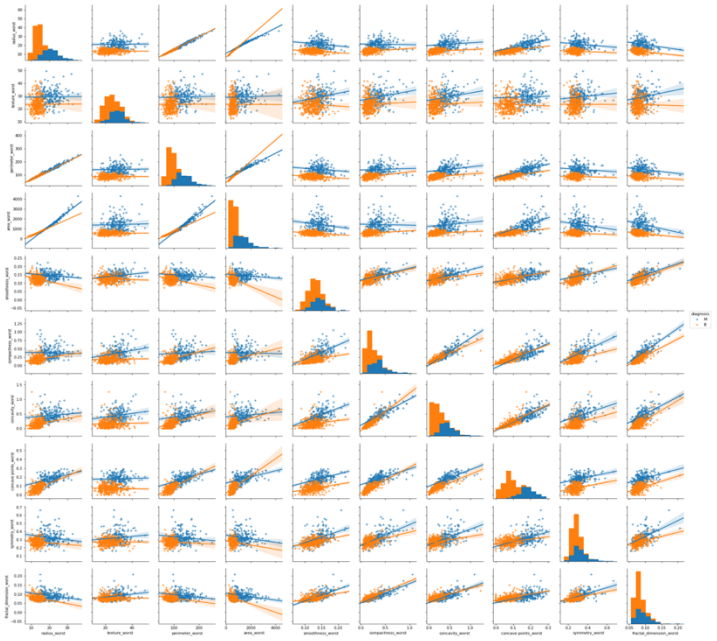

Figure 3: pair plot for 10 feature of worst group (blue = M and orange = B)

V. METHODS AND RESULTS

Based on the data analysis we made, we choose to use K nearest neighbors (KNN) and support vector machines (SVMs) because of the intensive overlap between the features both these to methods are the good way to solve that. Before we start implementing the classifier, we will need to standardization so we have mean equal zero and standard deviation equal one. Standardization is highly recommended since there are big difference between values and units. Python language will be the main language we will use, numpy, pandas, seaborn and sklearn packages will be used.

K-Nearest Neighbors (KNN)

For KNN we used two approaches, first we used all of features then fit it to KNeighborsClassifier model with k = 1 and parameter p equal 2 for Euclidian distance. After that test the performance of the classifier with confusion matrix and classification report that shows the precision, recall, f1-score and support of each classes and the average of them. Below text shows the parameter of the model and results of performance.

KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski', metric_params=None, n_jobs=1, **n_neighbors=1**, **p=2**, weights='uniform')

Confusion matrix		Classes	Precision	Recall	F1-score	Support
102	3	B	0.95	0.97	0.96	105
5	61	M	0.95	0.92	0.94	66
Avg / Total			0.95	0.95	0.95	171
Accuracy		0.95321637426900585				

In above table we can see that we have a good performance model. We can investigate more if we can improve this mode by studying what value of k can improve the model. Figure 4 shows the error rate with different values of k. This figure shows that we selected the best k value for minimum error rate.

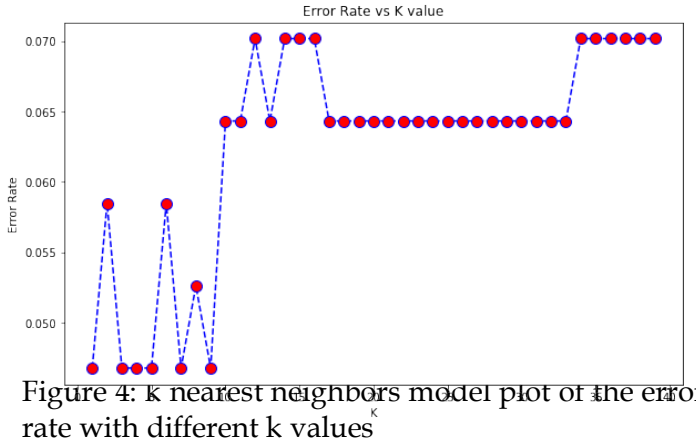

Figure 4: k nearest neighbors model plot of the error rate with different k values

Second approach is by applying the same process but on the three different group and then also investigate the better value of k and perform the performance measurement on it. We will start with mean group followed by standard error and then worst group.

Mean group with k =1:

Confusion matrix		Classes	Precision	Recall	F1-score	Support
69	9	B	0.96	0.91	0.94	105
4	62	M	0.87	00.94	0.91	66
Avg / Total			0.93	0.92	0.92	171
Accuracy		0.92397660818713445				

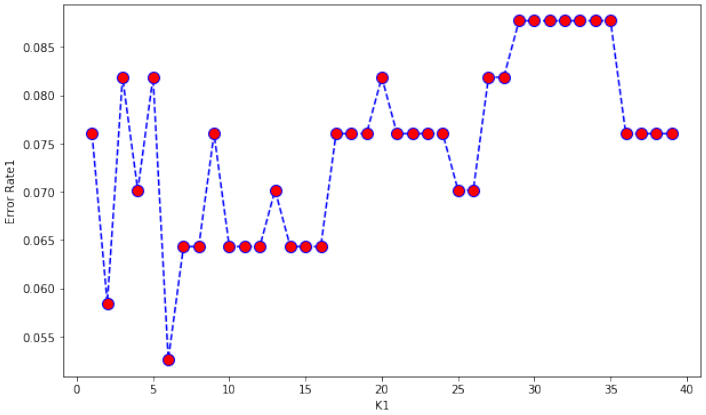

Figure 5: Mean group KNN model plot of the error rate with different k values

Mean group with best k =6:

Confusion matrix		Classes	Precision	Recall	F1-score	Support
102	3	B	0.94	0.97	0.96	105
6	60	M	0.95	0.91	0.93	66
Avg / Total			0.95	0.95	0.95	171
Accuracy		0.94736842105263153				

Standard error group with k =1:

Confusion matrix		Classes	Precision	Recall	F1-score	Support
93	12	B	0.89	0.89	0.89	105
11	55	M	0.82	0.83	0.83	66
Avg / Total			0.87	0.87	0.87	171
Accuracy		0.8654707602339176				

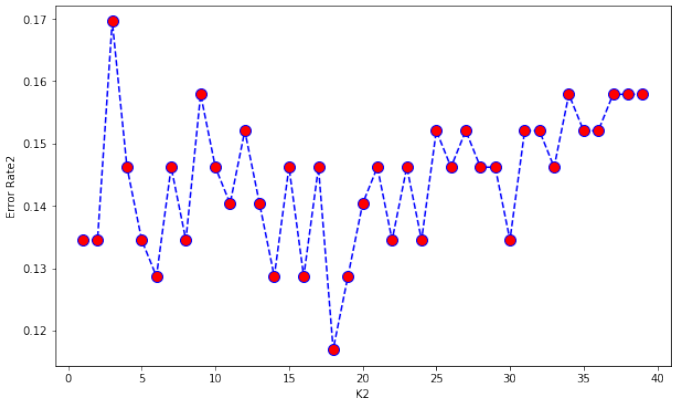

Figure 6: Standard error group KNN model plot of the error rate with different k values

Standard error group with best k =18:

Confusion matrix		Classes	Precision	Recall	F1-score	Support
93	12	B	0.89	0.89	0.89	105
6	60	M	0.95	0.83	0.83	66
Avg / Total			0.87	0.87	0.87	171
Accuracy		0.88304093567251463				

Worst group with k =1:

Confusion matrix		Classes	Precision	Recall	F1-score	Support
101	4	B	0.95	0.96	0.96	105
5	61	M	0.94	0.92	0.93	66
Avg / Total			0.95	0.92	0.95	171
Accuracy		0.94736842105263153				

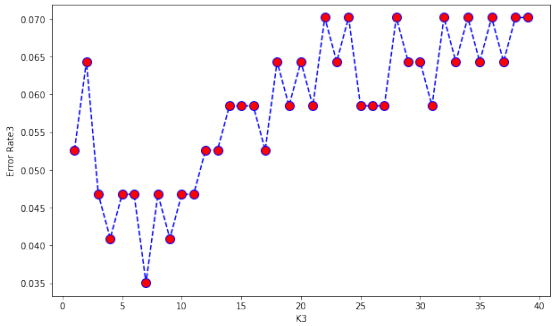

Figure 7: Worst group KNN model plot of the error rate with different k values

worst group with best k =7:

Confusion matrix		Classes	Precision	Recall	F1-score	Support
104	1	B	0.95	0.99	0.97	105
5	61	M	0.98	0.92	0.95	66
Avg / Total			0.97	0.96	0.96	171
Accuracy		0.96491228070175439				

To sum up all above results, we choose the best model based on the trade-off that we want higher accuracy for classification but also concern about the False Negative (FN) in the confusion matrix. FN is the misclassified of Malignant tumor as Benign tumor which we really care about, we do not want to classify disease as not dangerous state. Moreover, we can see if we used all features we can get FN equal 5 and accuracy of model equal 0.95 and we can achieve same FN value with better accuracy at 0.96 if we used best k value for worst group model.

Support Vectors Machines (SVMs)

We will apply the same two approaches on SVM where we use all feature and then apply it on different group. Over all, we will setup SVM model with default parameters and compute the performance measurement. Then we will see if we can improve it with grid search via cross validation on the parameters. The parameters that we will investigate are C and gamma. C control the cost of misclassification, large C value give you low bias and high variance and vice-versa as trad-off validation. Gamma is the free parameters in radial basis function and small gamma lead for gaussian of large variance and low bias. For the grid search we will investigate the performance with list of values, for C :[0.1,1,10,100,1000] , gamma :[1,0.1,0.01,0.001,0.0001] Below text shows the parameter of the first model and results of performance.

SVC(**C=1.0**, cache_size=200, class_weight=None, coe
f0=0.0,
decision_function_shape='ovr', degree=3, **gamma='a
uto'**, **kernel='rbf'**,
max_iter=-1, probability=False, random_state=None
, shrinking=True,
tol=0.001, verbose=False)

Confusion matrix		Classes	Precision	Recall	F1-score	Support
104	1	B	0.97	0.99	0.98	105
3	63	M	0.98	0.95	0.97	66
Avg / Total			0.98	0.98	0.98	171
Accuracy		0.97660818713450293				

All feature with grid search selecting C =100 and gamma = 0.0001:

Confusion matrix		Classes	Precision	Recall	F1-score	Support
105	0	B	0.95	1.00	0.97	105
6	60	M	1.00	0.91	0.95	66
Avg / Total			0.97	0.96	0.96	171
Accuracy		0.96491228070175439				

Mean group without GridSearchCV:

Confusion matrix		Classes	Precision	Recall	F1-score	Support
102	3	B	0.93	0.97	0.95	105
8	58	M	0.95	0.88	0.91	66
Avg / Total			0.94	0.94	0.94	171
Accuracy		0.93567251461988299				

Mean group with grid search selecting C =1 and gamma = 0.1:

Confusion matrix		Classes	Precision	Recall	F1-score	Support
102	3	B	0.93	0.97	0.95	105
8	58	M	0.95	0.88	0.91	66
Avg / Total			0.94	0.94	0.94	171
Accuracy		0.93567251461988299				

Standard error group without GridSearchCV:

Confusion matrix		Classes	Precision	Recall	F1-score	Support
96	9	B	0.86	0.91	0.89	105
15	51	M	0.85	0.77	0.81	66
Avg / Total			0.86	0.86	0.86	171
Accuracy		0.87134502923976609				

Standard error group with grid search selecting C =100 and gamma = 0.01:

Confusion matrix		Classes	Precision	Recall	F1-score	Support
95	10	B	0.89	0.90	0.90	105
12	54	M	0.84	0.82	0.83	66
Avg / Total			0.87	0.87	0.87	171
Accuracy		0.87134502923976609				

Worst group without GridSearchCV:

Confusion matrix		Classes	Precision	Recall	F1-score	Support
105	0	B	0.67	1.00	0.80	105
52	14	M	1.00	0.21	0.35	66
Avg / Total			0.80	0.70	0.63	171
Accuracy		0.88304093567251463				

Worst group with grid search selecting C =1000 and gamma = 1:

Confusion matrix		Classes	Precision	Recall	F1-score	Support
100	5	B	0.87	0.95	0.91	105
15	51	M	0.91	0.77	0.84	66
Avg / Total			0.89	0.88	0.88	171
Accuracy		0.88304093567251463				

As mention above, we more concern about the FN and accuracy over all. It seems from our results that the best performance was using all feature with default setup unlike the KNN methods. The highest accuracy was 0.97 close to the KNN with k = 7 on worst group. However, the SVM on all feature gave us the lowest FN = 3.

VI. COMPARISON OF DIFFERENT APPROACHES

Comparing the results, we got with one of the research paper that work on the same data. The paper applied Naïve Bayes, Random Forest, KNN, SVM and logistic regression []. The research achieves relatively similar results for 70% as train data and 30% as test data. Their KNN achieved accuracy of 96.4612% and SVM 97.4874 while we got 96.4912% for KNN and 97.6608 for SVM. The paper also tried different splitting percentages where the SVM achieved 98.2456% at 10% test and 90% train data [].

VII. CONCLUSION

Based on the data analysis we made, we decide to design classifier using k nearest neighbors (KNN) and support vectors machines (SVMs). We applied different approaches on KNN where we used all the feature and another trying the model on different group. In addition we investigated different value of k's. We achieved 96.491% accuracy and misclassification of malignant equal 5. For SVMs, we applied the same approached interim of features. Furthermore, we investigated different values for C (control cost of misclassification) and gamma (free parameters of radial basis function) parameters. We

achieved misclassification of malignant equal 3 and accuracy equal 97.661%. In future, we can use different techniques of ensemble either to select or fusion (hard and soft) for the classifier. Moreover, we can combine the principle component analysis (PCA) to KNN and SVMs classifiers. In addition, we can apply different splitting percentages on the datasets.

VIII. REFERENCES

[1]. "U.S. Breast Cancer Statistics." Breastcancer.org. Accessed May 05, 2018. http://www.breastcancer.org/symptoms/understand_bc/statistics.

[2]. "What Is Breast Cancer?" Breastcancer.org. Accessed May 05, 2018. http://www.breastcancer.org/symptoms/understand_bc/what_is_bc.

[3]. "Breast Cancer Risk and Risk Factors." Breastcancer.org. Accessed May 05, 2018. http://www.breastcancer.org/symptoms/understand_bc/risk.

[4]. "Fine Needle Aspiration Biopsy of the Breast." *American Cancer Society*, www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/breast-biopsy/fine-needle-aspiration-biopsy-of-the-breast.html.

[6]. Hook, Debra-Lynn. "When to Worry About Breast Lumps." Everyday Health. September 07, 2017. Accessed May 05, 2018. https://www.everydayhealth.com/womens-health/when-to-worry-about-breast-lumps.aspx.

[7]. "Different Kinds of Breast Lumps." Cancer Center. May 20, 2015. Accessed May 05, 2018. https://cancer.stonybrookmedicine.edu/breast-cancer-team/patients/bse/breastlumps.

[8]. H. Jouni, M. Issa, A. Harb, G. Jacquemod and Y. Leduc, "Neural Network architecture for breast cancer detection and classification," *2016 IEEE International Multidisciplinary Conference on Engineering Technology (IMCET)*, Beirut, 2016, pp. 37-41.

[9]. C. Shahnaz, J. Hossain, S. A. Fattah, S. Ghosh and A. I. Khan, "Efficient approaches for accuracy improvement of breast cancer classification using wisconsin database," *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, Dhaka, 2017, pp. 792-797.

[10]. H. Rajaguru and S. Kumar Prabhakar, "Bayesian linear discriminant analysis for breast cancer classification," *2017 2nd International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, 2017, pp. 266-269.

[11]. K. Khasburrahman, A. Wibowo, I. Waspada, H. B. Hashim and W. Jatmiko, "Comparison of diagnostics set and feature selection for breast cancer classification based on microRNA expression," *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)*, Semarang, 2017, pp. 165-170.