Breast Cancer Prediction

# **Bhushan Suryawanshi** [**DSC680-T302 Applied Data Science (2215-1)**](https://cyberactive.bellevue.edu/webapps/blackboard/execute/courseMain?course_id=_500232_1)

# Bellevue University

**Abstract:**

Lifetime breast cancer risk in U.S. women is 1 in 8 that is about 13%. In 2021 it is expected that the US will add 281,550 cases of breast cancer in women. If it can be detected and treated early, studies show that these statistics can be improved. The purpose of this project was to determine if machines can accurately assist doctors, particularly breast cancer specialists, in making an accurate predictive diagnosis relating to breast cancer.   
In this study, the data provided by the Diagnostic Wisconsin Breast Cancer database is used to analyze and run through four different predictive models. In order to accurately determine if the predictive model could make predictions, multiple characteristics were evaluated to determine their value in our formula. Since this is a classification prediction problem, the following models were selected for evaluation: Random Forest Classifier, Logistic Regression CV, K-Nearest Neighbor Classifier, and Support Vector Machines (SVM) with Support Vector Classifier. A simple function was created for displaying each model’s metrics. Each model was fit and tuned using the training data set, and a prediction was made and evaluated using the testing data. Models have been evaluated in Python and R. Each model was able to accurately predict a correct diagnosis of a cancerous tumor over 90% of the time. These results are very exciting and help in breast cancer diagnosis.

**Keywords**: K-Nearest Neighbor(KNN), Support Vector Machine(SVM), Logistic regression, Random Forest, Classification, Cancer, Machine Learning.

**Introduction:**

Cancer is a disease that attacks thousands of people annually. The ability to quickly and accurately diagnose this disease greatly increases one’s ability to overcome and survive this affliction.

Currently, breast cancer is diagnosed by completing one or more the following:

1. **Breast ultrasound**: This uses ultrasound waves to get images of the internal structure of the breast.
2. **Diagnostic mammogram**: It is the x-ray of the breast which reveals the cancer tumor formation.
3. **Magnetic resonance imaging (MRI)**: These machines use high magnetic radiations to get the scans of the breast which also reveals the cancer tumor in the soft tissue part of the breast that cannot be found with other scans.
4. **Biopsy**: This procedure uses small tissue of the breast for laboratory testing.

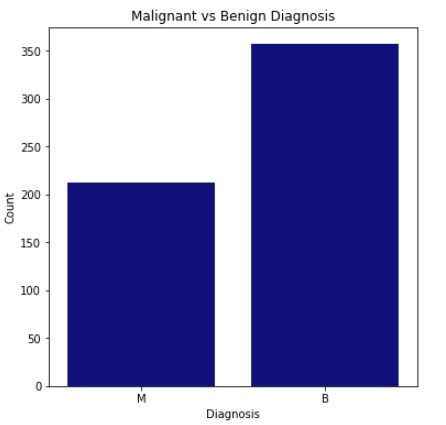
Analyzing and using laboratory reports can also help in creating a tool that can help early detection of cancer.   
The "Diagnostic Wisconsin Breast Cancer Database" is a publicly available data set from the UCI machine learning repository. Breast cancer testing has a process called Fine Needle Aspirate (FNA). This process uses breast mass collected using a hollow needle attached to a syringe to get a small amount of tissue. Which is used to run laboratory tests. The data is collected from these tests and made available for machine learning.

Machine learning has helped to solve the problems of many industries. The Healthcare industry is one of the leading users of technological advancement. The huge availability of health data has helped to create very advanced models. This technological change not only helping patients but also the clinical staff. Health experts who used to analyze the reports manually can now be ready for them by an algorithm. It is helping to speed up the detection of health complications in the early stages. This study is one such effort to analyze and predict breast cancer in an early stage. As it has been seen early cancer detection has helped to improve chances of recovery. With other image classification technologies, this method will be an additional tool in the arsenal of healthcare experts to detect breast cancer.   
The goal of this project is to examine the results of multiple features being utilized in providing a predictive value between benign and malignant breast tumors.

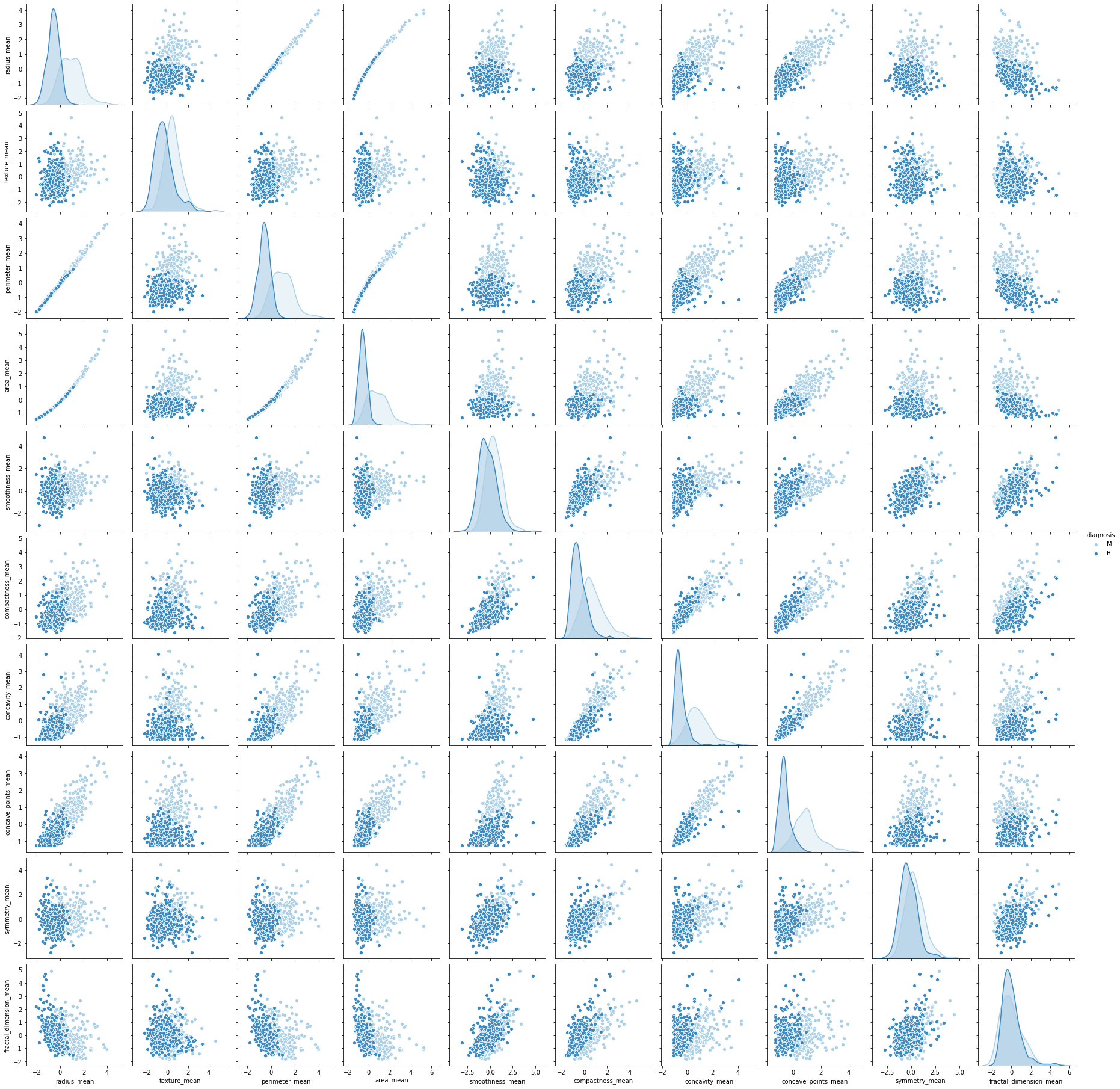
**Approach:**

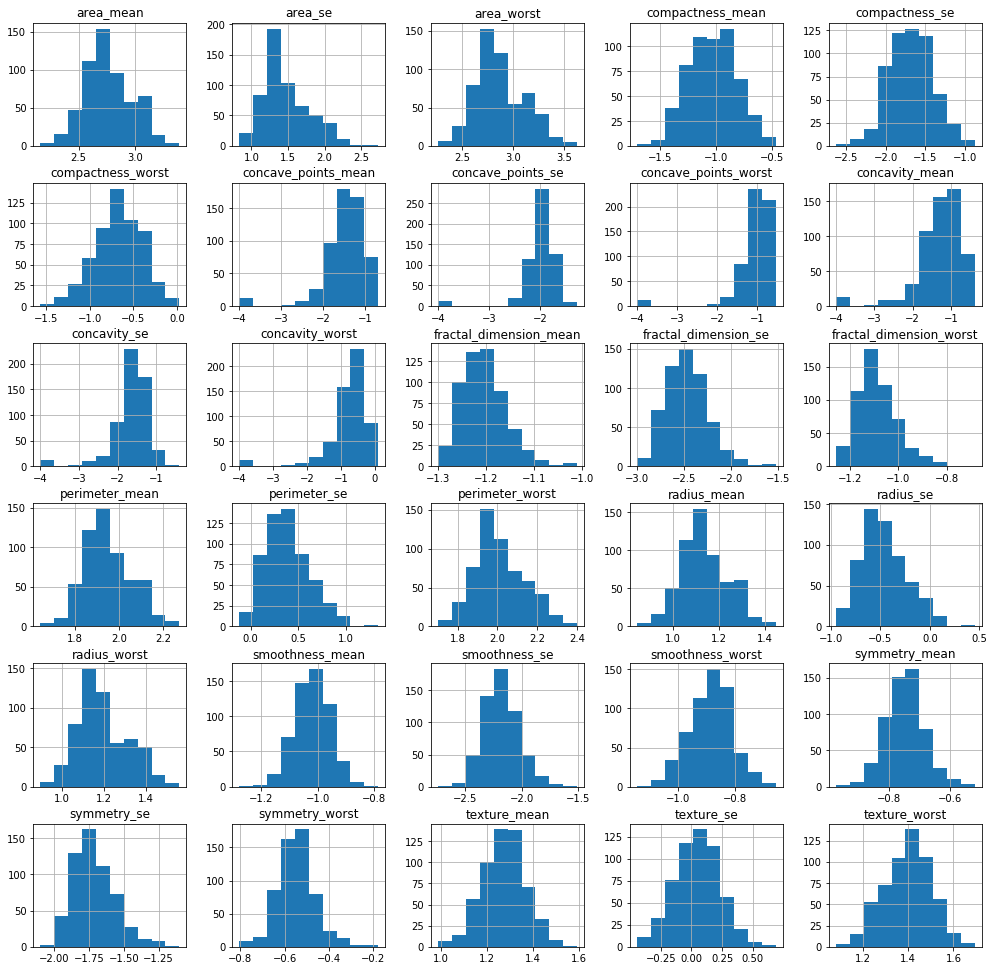
**Data Preparation:** For data preparation, Python was utilized to read in and clean the initial dataset. The work required was relatively less than expected to get the working dataset. There was an unnamed column in the dataset. Upon the removal of the unnamed column, analyzed dataset for null and empty fields. The analysis showed that there are no NULL or empty field values. Also, the duplicate check was passed by the dataset with zero duplicate values. The duplicate check was performed using the “id” column.

**Exploratory Analysis:** Exploring the data has led to some basic and interesting discoveries. The data consists of 569 records, with 212 as malignant and 357 diagnosed as benign, 37.26% and 62.74% respectively.



The dataset has features such as mean, standard error, worst (largest of standard error) were normalized and plotted in box plots and pair plots to analyze statistical summaries and relationships between the features and diagnosis. The box plot showed a high number of outliers for both malignant and benign diagnosed growths. Since the dataset is a smaller sample size, 529 records, the outliers remained in the analysis. Also, outliers may be because of abnormal growth of tissue and hence it is difficult to truly identify each of the outliers as a true outlier. With the pair plot, we are able to visualize initial patterns of how the features and diagnosis relate to one another.

Each feature’s distribution was plotted, it became clear that most were skewed to the right. All features were thus transformed using log10 (with 0.001 being added to account for zero values in the features). This allowed most features to form normal distributions, allowing for better analysis and avoiding future overfitting. The following graph shows the distribution of data after log transformation.



**Feature Selection:** Feature selection was the most important and challenging portion of the analysis. With the data set being relatively small, 529 records, and many of the features having high correlations, it was vital to identify the most important features, but also not to eliminate those with minor significance. This was needed to preserve the integrity of the data, but also to provide a broader scope in the model’s ability to accurately make consistent, valid predictions. At first, a Lasso linear cross-validation model was executed. LassoCV shrinks the coefficients to zero for less relevant features. This helped to reduce the number of features from 30 to 23. After this reduction was completed, a correlation heatmap was constructed. It has been seen that many features still have high correlation rates. PCA while reducing the features does not account for class labels. It preserves only the maximum variance. This is not optimal for classification problems. This is the reason LassoCV was used for feature selection. In order to account for high multicollinearity, an Ordinary Least Squares Regression model was used. Each feature with a p-value greater than 0.05 is eliminated maintaining an R-squared of more than 0.750. After multiple passes, the final dataset was constructed with 11 features being selected.

**Model building*:*** The dataset was randomly split into 80% training and 20% testing data sets. Since this is a classification prediction problem, the following models were selected for evaluation: Random Forest Classifier, Logistic Regression CV, K-Nearest Neighbor Classifier, and Support Vector Machines (SVM) with Support Vector Classifier. For metrics visualization, a simple function was created that can be used with each model. Each model was fit and tuned using the training data set, and a prediction was made and evaluated using the training data. Models were evaluated in both Python and R.

Results: The results from each model are shared below.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model Metrics | Train Accuracy Mean (CV=10) | Test Accuracy Mean (CV=10) | AUC Score | Precision | Recall | F1 Score | Accuracy |
| Random Forest | 0.9273 | 0.9303 | 0.9672 | 0.96 | 0.97 | 0.96 | 0.96 |
| Logistic Regression | 0.9714 | 0.9561 | 0.9415 | 0.94 | 0.94 | 0.94 | 0.95 |
| KNN (n=3) | 0.9472 | 0.9561 | 0.9605 | 0.95 | 0.96 | 0.95 | 0.96 |
| SVM | 0.9582 | 0.9644 | 0.9477 | 0.94 | 0.95 | 0.94 | 0.95 |

After tuning we can see that each of the models performed extremely well, with all the metrics showing good numbers. Logistic Regression did return a slightly lower test accuracy score than the training model, but the difference is relatively small. Additional data points and features might help to improve logistic regression, keeping it for future work. While looking at each model performance it can be observed that Random Forest Classifier did a little better in each category compared to the other models. This can be a model of choice based on the current study accomplished. Random forest algorithms have been worked well with classification problems as they can handle both discrete and continuous variables well. Another reason random forest works better is that they work by an ensemble of multiple decision trees during training.

**Conclusion**

As we have seen from the results provided, all models performed much better than expected levels. The data sample used in this study was very small that might be questionable in terms of generalizing findings for the larger setup. Hence, before productionizing any of these models it is recommended to train and evaluate the model on a larger dataset with the same constraint to confirm the findings from this study. With the current number, the Random Forest classifier is best to use as the first model to be used in production. Overall, the study highlights that predictive analytics can help and act as a tool in the early detection of breast cancer. Other areas of medical study can be benefited from these findings. Relying completely on predictive models may not be ideal for understanding the involvement of human life. However, by working in collaboration with experts like oncologists and radiologists, the data scientist can help in the diagnostic study.

**References -**

1. American Cancer Society. (2020, January). How Common Is Breast Cancer? - <https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html>.

2. American Cancer Society. Cancer Facts & Figures 2020. <https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2020/cancer-facts-and-figures-2020.pdf>.

3. National Cancer Institute. (2018, January). BRCA Mutations: Cancer Risk and Genetic Testing. <https://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet>.

4. Ak MF. A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications. Healthcare (Basel). 2020;8(2):111. Published 2020 Apr 26. doi:10.3390/healthcare8020111 [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7349542/]

5. Raghupathi W., Raghupathi V. Big data analytics in healthcare: Promise and potential. Health Inf. Sci. Syst. 2014;2:3. doi: 10.1186/2047-2501-2-3. [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4341817/]

6. Petri I., Kubicki S., Rezgui Y., Guerriero A., Li H. Optimizing energy efficiency in operating built environment assets through building information modeling: A case study. Energies. 2017;10:1167. doi: 10.3390/en10081167. [https://www.mdpi.com/1996-1073/10/8/1167#cite]

7. Jothi N., Rashid N.A., Husain W. Data mining in healthcare—A review. Procedia Comput. Sci. 2015;72:306–313. doi: 10.1016/j.procs.2015.12.145. [https://www.sciencedirect.com/science/article/pii/S1877050915036066]

8. Dai H., Cheng Z., Baii J. Breast cancer cell line classification and its relevance with breast tumor subtyping. J. Cancer. 2017;8:3131–3141. doi: 10.7150/jca.18457. [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5665029/]

9. Rizki A, Weaver VM, Lee SY, Rozenberg GI, Chin K, Myers CA. et al. A Human Breast Cell Model of Preinvasive to Invasive Transition. Cancer research. 2008;68:1378–87. [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2662716/]

10. Kurebayashi J, Otsuki T, Kurosumi M, Yamamoto S, Tanaka K, Mochizuki M. et al. Isolation and characterization of a new human breast cancer cell line, KPL-4, expressing the Erb B family receptors and interleukin-6. British Journal of Cancer. 1999;79:707–17. [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2362677/]