**PREDICTING CANCER USING MACHINE LEARNING**

## Motivation

Inasive ductal carcinoma (IDC) is - with ~ 80 % of cases - one of the most common types of breast cancer. It's malicious and able to form metastases which makes it especially dangerous. Often a biopsy is done to remove small tissue samples. Then a pathologist has to decide whether a patient has IDC, another type of breast cancer or is healthy. In addition, sick cells need to be located to find out how advanced the disease is and which grade should be assigned. This has to be done manually and is a time-consuming process. Furthermore, the decision depends on the expertise of the pathologist and his or her equipment. Therefore machine learning could be of great help to automatically detect and speed up the process. This way one would be able to overcome the dependence on the pathologist which would be especially useful in regions where no experts are available.

## 

## Our goal

As we started with this analysis we asked ourselves if we would be able to improve the results that were presented in 2014 in the paper [Automatic detection of invasive ductal carcinoma in whole slide images with Convolutional Neural Networks](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.725.4294&rep=rep1&type=pdf) by professor [Anant Madabhushi](https://case.edu/medicine/ccir/faculty/anant-madabhushi). Many years have passed since then and it's very likely that all methods used in the paper have already been changed, so we are trying to improve the results

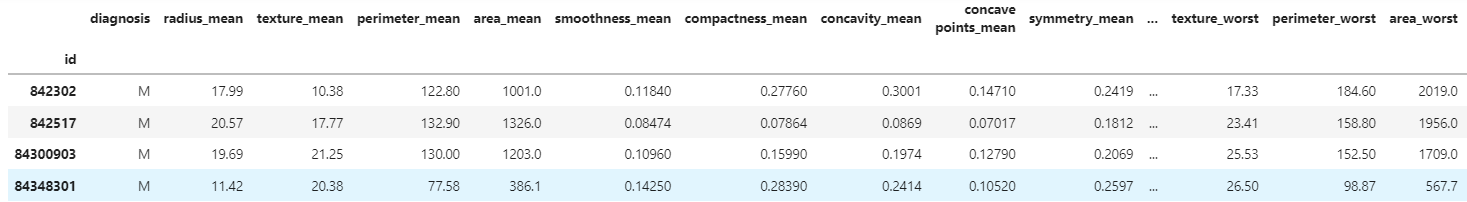
DATA

The data is extracted from [UCI ML repository](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29).

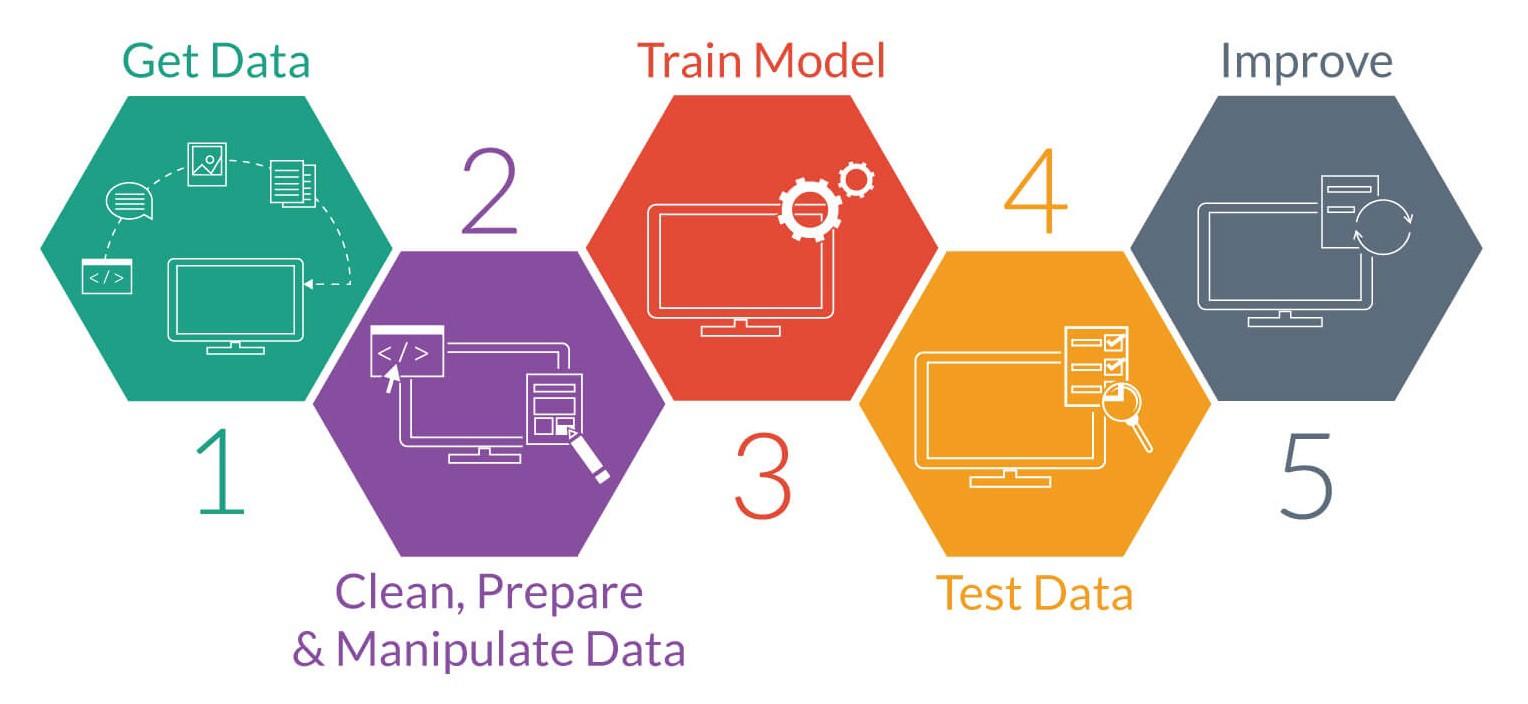
Ten real-valued features are computed for each cell nucleus:

* **radius**: distances from the center to points on the perimeter
* **texture**: standard deviation of gray-scale values
* **perimeter**
* **area**
* **smoothness**: local variation in radius lengths
* **compactness**: perimeter^2 / area - 1.0
* **concavity**: severity of concave portions of the contour
* **concave points**: number of concave portions of the contour
* **symmetry**
* **fractal dimension**: "coastline approximation" - 1

The sample data looks like this



Machine Learning Suff:



By using the data first we want to **cleaning the data** (missing elements, unwanted data, Normalizing the data) which brings a big impact on the results

The next step is to develop our model. We will start by first splitting our dataset into two parts; one as a training set for the model, and the other as a test set to validate the predictions that the model will make. We want to make sure that our model truly has predictive power and is able to accurately label unseen data. We will set the test size to 0.3; i.e., 70% of the data will be assigned to the training set, and the remaining 30% will be used as a test set. In order to obtain consistent results, we will set the random state parameter to a value of 40.

By using the LOGISTIC REGRESSION algorithm we have to train our model. logistic regression

The value ranges from 0 to 1, something that can take two values such as true/false, in our case (M-malignant, B-begins)

**Malignant is a term used to describe cancer**

**Begins is a term used to describe tumors that aren't cancer**

**PREDICTING THE MODEL AND EVALUATING**

We have successfully developed a logistic regression model. This is the key feature of a logistic regression model. However, for us to evaluate whether the predictions are accurate, the predictions must be encoded so that each instance can be compared directly with the labels in the test data. In other words, instead of numbers between 0 or 1, the predictions should show "M" or "B", denoting malignant and benign respectively. In our model, a probability of 1 corresponds to the "Benign" class, whereas a probability of 0 corresponds to the "Malignant" class.

**COST ANALYSIS:**

The average biopsy Test cost range from **5000 Rs to 8000 Rs** and even more in private hospitals

But by using our model we can find if it is cancer or not while sitting at the home. Moreover, it gives doctors more clarification and confidence about the tumor and we call also build a user-friendly interface between doctors and patients by consulting online. Also, the error in false prediction of cancer tumors is less than the error in false prediction begins.

**CONCLUSION**

**By using this project we can find cancer early and easy and save millions of people's life which brings a sustainable future to our country.**

**Thank you**