Domain Background In this project I am going to address the medical problem of the breast tumor classification [4]. When a patient is suspected to have a breast tumor, a complete diagnosis is required in order to investigate whether the tumor is benign or a cancer (malign). Many models have already been published, such as Gail (1989) [5], Claus (1991) [2],Claus (1994) [3], Antoniou (2004) [1], Jonker (2003) [6], Van Asperen (2004) [9] and Tyrer (2004) [8].

Using a dataset of real data of medical analysis of patients, I am going to apply statistical prediction models to predict whether a breast cancer is malign or benign. The dataset has been published by the University of California, Irvine on Kaggle and it has already been analyzed by some users on Kaggle and also some researchers [7] with great results. Alessandro Peretti and Francesco Amenta used logistic regression with CUDA parallel programming for their analysis. The first thing I notice is that the dataset is composed by 569 instances, which is a fairly slow dataset to justify the use of GPU parallel computing. Furthermore, they used logistic regression which is an appropriate model because the output is supposed to be a percentage of the risk. In fact, the number of wrong classifications performed by their model is still pretty high (9% of wrong classifications). Some Kaggle users have used a support vector machine to classify the instances and they obtained high values of accuracy (95-98%).

Problem Statement I am going to develop a machine learning model that, given a small dataset of patients affected by breast tumor with their diagnostic data, will predict whether the tumor is a cancer or not. This is clearly a classification problem because the output ideally should be yes or no. If we look at the "diagnosis" feature in the dataset we notice that it can be "M" as malignant or "B" as benign. We will change those labels in 1 for malignant and 0 for benign. This will be the target feature to be used in the training phase and to verify the prediction in the test set. However, most of the researchers who worked on this research have treated this as a regression problem. The reason is that they want to provide to the doctor a tool capable of quantifying the probability that such a tumor is a cancer or not. Potentially, if this research will be used to develop a real medical tool, the machine learning should output the probability. However, for the purposes of this research, I will treat this as a classification problem.

In this field, false positives are as dangerous as true negatives: imagine to treat a benign tumor as a cancer. The patient might die under the collateral effects of a treatment that wasn't really needed. On the other end, if we treat a cancer as a benign tumor the patient will die because the treatment

is not adequate. Given the size of the dataset and the importance of avoiding misclassifications, I will use boosting algorithms in order to avoid any type of overfitting issue.

Datasets and Inputs The dataset has been downloaded from Kaggle and it is composed by a single CSV file. The dataset contains 569 instances characterized by 30 features in addition to the classification column that has been donated in 1995. [7] provides a Pearson Correlation Coefficient of the dataset which, the authors says, demonstrates that the features are independent and this makes inefficient the application of any reduction techniques, such as Principal Component Analysis. Looking at the graph shown in their article, I can see there is some dependence among some of the features. Even though the graph is scattered, I think applying PCA might provide some good results. In my research I will calculate the Pearson correlation coefficient and I will try to reduce the dataset with PCA.

I also noticed that some of the features have a range of values very high. It is appropriate to apply scaling techniques to the dataset before applying machine learning algorithms.

Solution Statement I will start with a deep analysis of the dataset. I will identifying possible outliers and assessing how those are catalogued by a support vector machine. This will help me understanding if boosting techniques might perform better than a simple SVM. I will verify the results of my analysis applying Ada Boosting to the dataset and measure the results. If the error rate is due to the variance of the data and overfitting issues due to the limited size of the dataset, I might be able to improve those performances using a boosting technique in my model.

Benchmark Model In this field, false positives are as dangerous as true negatives: imagine to treat a benign tumor as a cancer. The patient might die under the collateral effects of a treatment that wasn't really needed. On the other end, if we treat a cancer as a benign tumor the patient will die because the treatment is not adequate. For this reason I will set the cutoff to 0.5, meaning that a probability lower than 0.5 will be considered benign, while a probablity equal or higher than 0.5 will be classified as cancer. To start, I will keep it a binary classification problem, because this gives me access to few additional methods inside sklearn.

The performances of my boosting models are measurable through the comparison of the outputs with the results column in the dataset. I will also

compare the F1 value with the SVM and the F1 value obtained with boosting in order to understand if a different algorithm can improve the performance.

Evaluation Metrics Considering that this is a binary classification problem, the most appropriate metric to use is F1. This metric is more appropriate than the simple accuracy because it takes into account false positives and false negatives in proporsion with the correct predictions.

Project Design My project will begin with a deep analysis of the dataset. I will calculate the mean, median and variance of the data. I will use the percentile calculation to detect possible outliers in the dataset and use a SVM fitted on the training dataset to make predictions for those instances. This will give me an understanding whether the error rate of SVM might be due to outliers instances or not. I will reason on the data obtained and produce a statement describing the expected results by the application of boosting techniques.

In the next phase I will apply SVM and Ada Boosting classifiers and calculate the respective F1 metric. The boosting classifier will be used with a support vector machine, a classification tree and a K Neightbors classifier.

It will follow a comparison table with the F1 values obtained by each classifier.

References

[1] AC Antoniou, AP Cunningham, J Peto, DG Evans, F Lalloo, SA Narod, HA Risch, JE Eyfjord, JL Hopper, MC Southey, et al. The boadicea model of genetic susceptibility to breast and ovarian cancers: updates and extensions. *British journal of cancer*, 98(8):1457–1466, 2008.

[2] Elisabeth B Claus, N Risch, and W Douglas Thompson. Genetic analysis of breast cancer in the cancer and steroid hormone study. *American journal of human genetics*, 48(2):232, 1991.

[3] Elizabeth B Claus, Neil Risch, and W Douglas Thompson. Autosomal dominant inheritance of early-onset breast cancer. implications for risk prediction. *Cancer*, 73(3):643–651, 1994.

[4] Susan Ely and Anna N Vioral. Breast cancer overview. *Plastic Surgical Nursing*, 27(3):128–133, 2007.

[5] Mitchell H Gail, Louise A Brinton, David P Byar, Donald K Corle, Sylvan B Green, Catherine Schairer, and John J Mulvihill. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute*, 81(24):1879–1886, 1989.

[6] MA Jonker, CE Jacobi, WE Hoogendoorn, NJD Nagelkerke, Geertruida H De Bock, and Johannes C Van Houwelingen. Modeling familial clustered breast cancer using published data. *Cancer Epidemiology Biomarkers & Prevention*, 12(12):1479–1485, 2003.

[7] A Peretti and F Amenta. Breast cancer prediction by logistic regression with cuda parallel programming support. *Breast Can Curr Res*, 1(111):2, 2016.

[8] Jonathan Tyrer, Stephen W Duffy, and Jack Cuzick. A breast cancer prediction model incorporating familial and personal risk factors. *Statistics in medicine*, 23(7):1111–1130, 2004.

[9] Christi J van Asperen, MA Jonker, CE Jacobi, JEM van Diemen-Homan, E Bakker, MH Breuning, JC Van Houwelingen, and GH De Bock. Risk estimation for healthy women from breast cancer families new insights and new strategies. *Cancer Epidemiology Biomarkers & Prevention*, 13(1):87–93, 2004.