ARTIFICIAL INTELLIGENCE

FINAL DOCUMENTATION

Sahana Nalla

University of New Haven

00658650

Machine learning is simply making healthcare smarter. This powerful subset of artificial intelligence may be familiar to many in use cases such as speech recognition used by voice assistants, and in creating personalized online shopping experiences through its ability to learn associations. However, machine learning has demonstrated truly life-impacting potential in healthcare – particularly in the area of medical diagnosis.

To demonstrate how machine learning and deep learning are able to provide a medical diagnosis, I’ll walk you through a step-by-step example of how the technology can be used to detect and diagnose breast cancer using a publicly available data set.

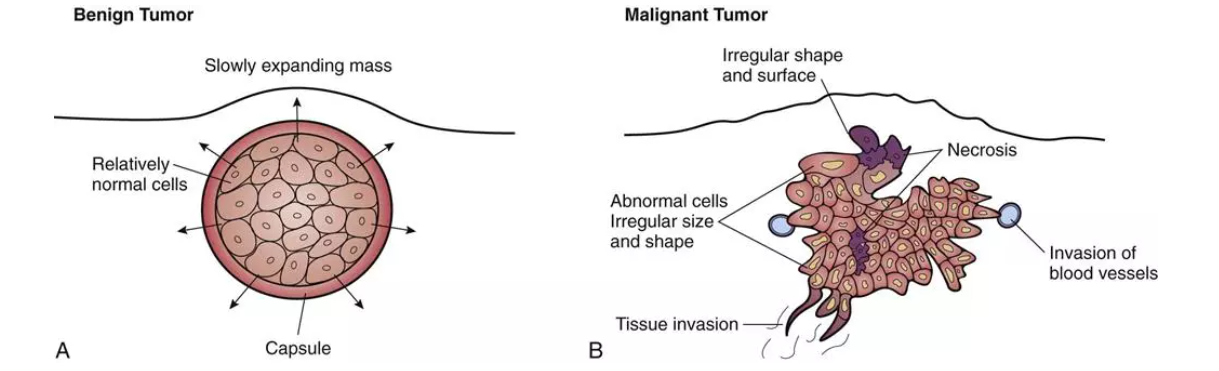
**Challenges of Applying Machine Learning in Healthcare**

There are several obstacles impeding faster integration of machine learning in healthcare today. One of the biggest challenges is the ability to obtain patient data sets which have the necessary size and quality of samples needed to train state-of-the-art machine learning models. Since patient data is protected by strict privacy and security rules, the data is not easy to collect, share and distribute. Furthermore, there are challenges with the format and quality of data which usually require significant effort to clean and prepare for machine learning analyses.

As machine learning and data science are starting to be adopted as a tool in healthcare applications, the industry is slowly pushing the boundaries on what it can do. Its primary function will most likely involve data analysis based on the fact that each patient generates large volumes of health data such as X-ray results, vaccinations, blood samples, vital signs, DNA sequences, current medications, other past medical history, and much more.

**Using Machine Learning to Detect and Diagnose Breast Cancer**

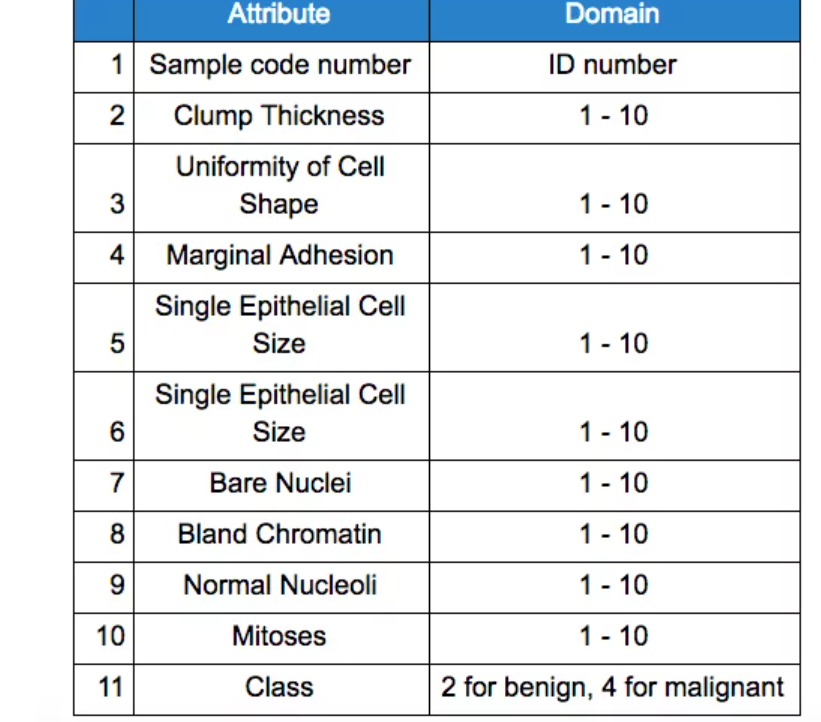
One application of machine learning in a healthcare context is digital diagnosis. ML can detect patterns of certain diseases within patient electronic healthcare records and inform clinicians of any anomalies. In this sense, the artificial intelligence technique can be compared to a second pair of eyes that can evaluate patient health based on the knowledge extracted from big data sets by summarizing millions of observations of diseases that a patient could possibly have. To illustrate just how useful machine learning as a medical diagnosis tool can be, I examined its use in breast cancer detection using a publicly available Breast Cancer Wisconsin (Diagnostic) Data Set.



*Characteristics of benign and malignant tumors*.

This data set consists of several instances of tumors. Tumors can either be benign (non-cancerous) or malignant (cancerous). Benign tumors grow locally and do not spread. As a result, they are not considered cancerous. However, they can still pose a danger, especially if they press against vital organs like the brain. Malignant tumors, in contrast, have the ability to spread and invade other tissues. This process, known as metastasis, is a key feature of cancer. There are many different types of malignancy-based tumors as well as locations that this type of cancer tumor can originate, as described in the data set specification.

The breast cancer data set consists of 699 tumor samples where 458 (65.5%) are benign (non-cancer) tumors and 241 (34.5%) malignant (cancer) tumors. Instances in the data set have the following attributes:



Solving a problem with machine learning often involves many iterative experiments meant to find the best model for solving the problem by further tuning the model. Given that there are many machine learning algorithms and different neural network architectures, a researcher (based on his/her experience, knowledge and trusting his/her intuition) will select the most promising model to set up the first experiment.

In our example, given the relatively small sizes of data sets, my intuition was to start modeling using traditional machine learning algorithms (e.g. SVM, KNN etc.) and shallow neural networks. To demonstrate some initial results using machine learning to diagnose breast cancer, the following set of metrics are used: ROC curve ≈ 0.99, Precision-Recall curve ≈ 0.99, and F1 ≈ 0.97.

**Setting up the Diagnosis Model**

**Step 1: Dividing the Data Set**

In order to get started modeling, the data set was split into two parts:

1. Train set (70%), for choosing and validating models, and
2. Test set (30%), hold out data on which we will see how well models are able to generalize on unseen data.

**Step 2: Defining the Metrics**

Next, we need to define the key metrics to measure the efficiency of the models. In order to describe the classifiers’ performance in the digital diagnoses problem, we have four basic characteristics (numbers) based on which we can define derivative measurement metrics. These four numbers are:

1. TP (True Positive) – number of correctly classified patients who have the disease,
2. TN (True Negative) – number of correctly classified patients who are healthy,
3. FP (False Positive) – number of misclassified patients who are healthy,
4. FN (False Negative) – number of misclassified patients who have the disease.

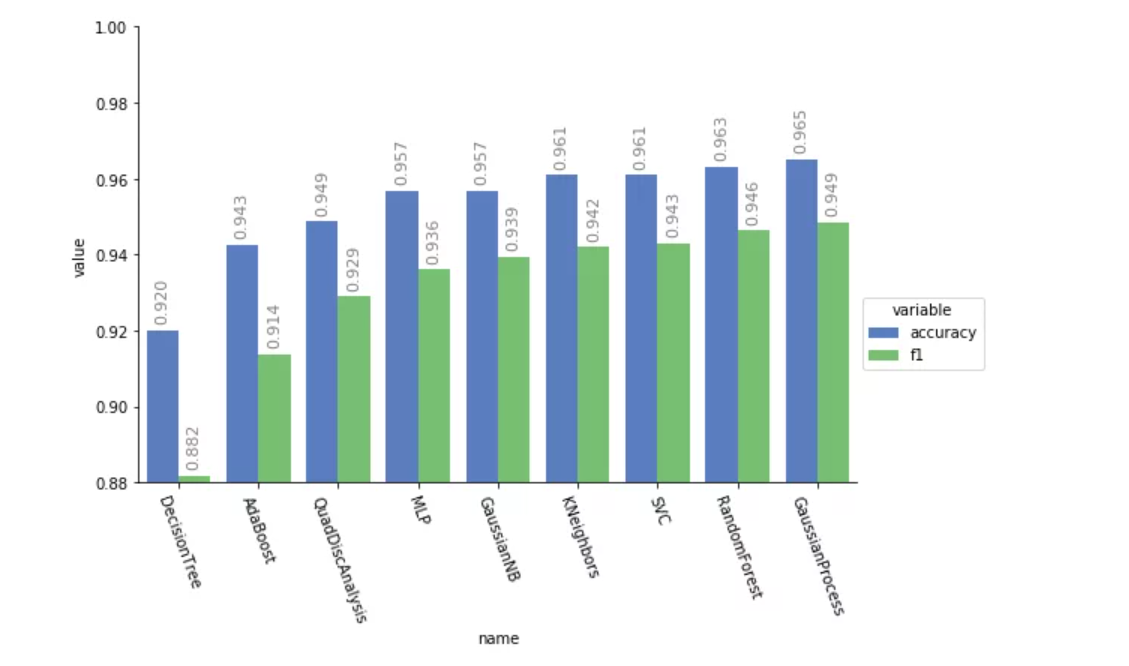
Based on these numbers we define the metrics as follows:

* Accuracy – ratio of correctly classified patients to the total number of patients (Accuracy = (TP+TN)/(TP+FP+FN+TN))
* Precision – ratio of correctly classified patients with the disease to the total patients classified as having the disease. The intuition behind precision is how many patients classified as having disease truly have the disease (Precision = TP/TP+FP).
* Recall – ratio of correctly classified diseased patients to patients who have the disease. The intuition behind recall is how many patients who have disease classified as having the disease. (Recall = TP/TP+FN).

**Step 3: Evaluating the Models**

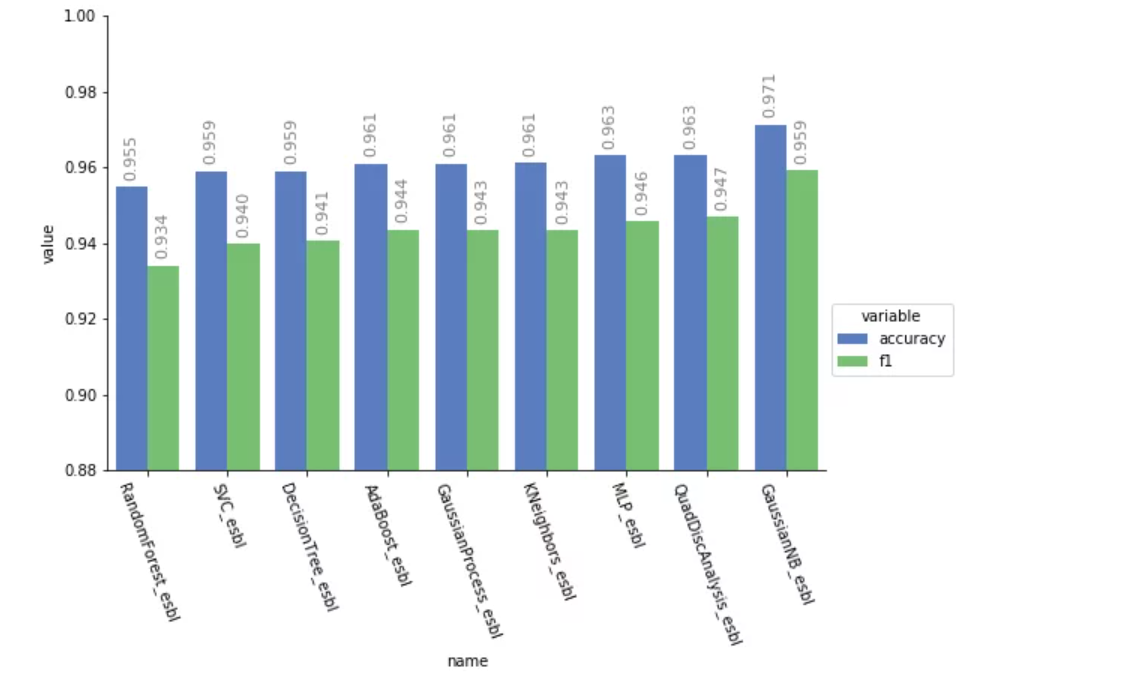
The next step involves using precision and recall metrics to evaluate the models. For the sake of simplifying the comparison of various models, we will use the harmonic mean of precision and recall which is called an F1 score (F1 Score = 2\*(Recall \* Precision) / (Recall + Precision)).

After experimenting with different algorithms, the mean F1 scores, in cross-validation, gained by each classifier is presented below. Given that accuracy is considered the most intuitive measure, it has also been plotted on the graph.



*Cross-validation scores of Machine Learning models.*

As you can see from the graph, the classifiers are showing pretty good results in terms of being able to better distinguish patients who have cancer versus those who are healthy by reaching 0.94 F1 scores. Where the best value for F1 is 1, and the worst value is 0. In order to gain higher scores, ensembles of these models were created by using bagging techniques.

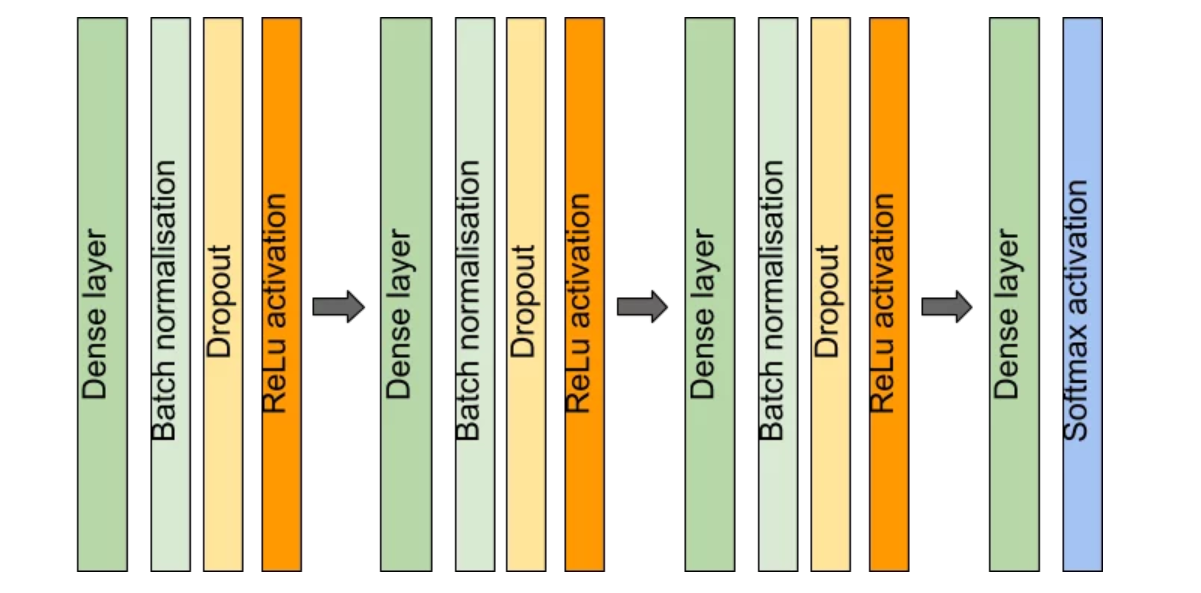


*Cross-validation scores of the ensemble Machine Learning models.*

As shown in the graph, the ensembles of models performed even better by reaching 0.95 F1 scores.

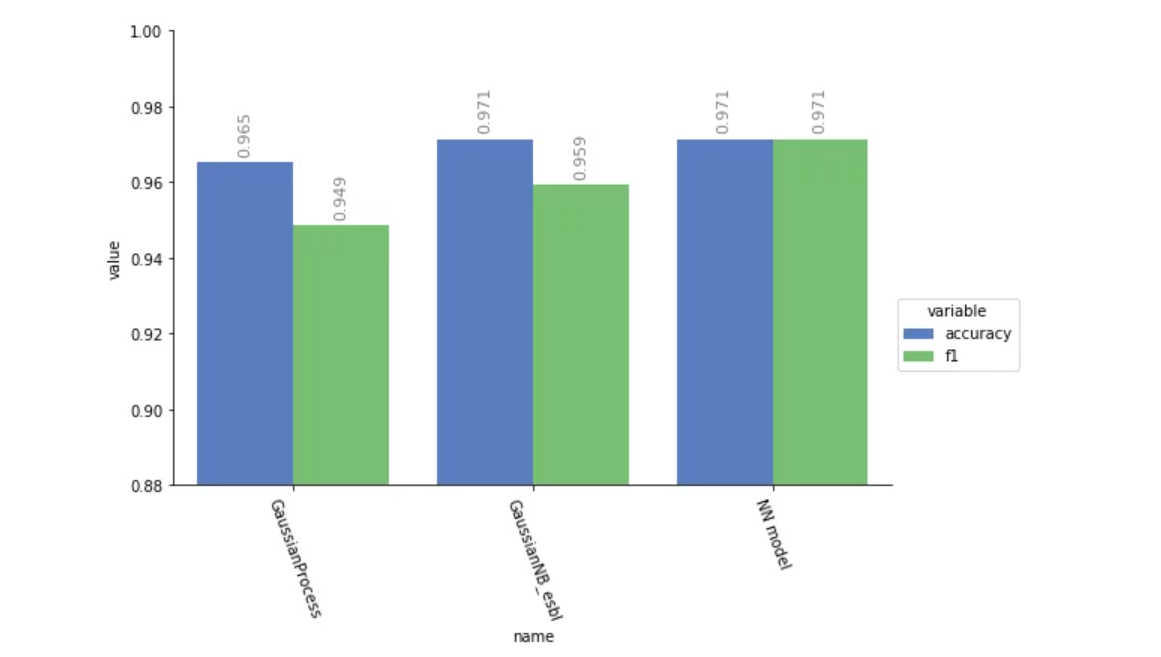
**Step 4: Creating a Neural Network Model**

In addition to the aforementioned diagnostic models, a Neural Network model was created and tuned using the architecture shown below.



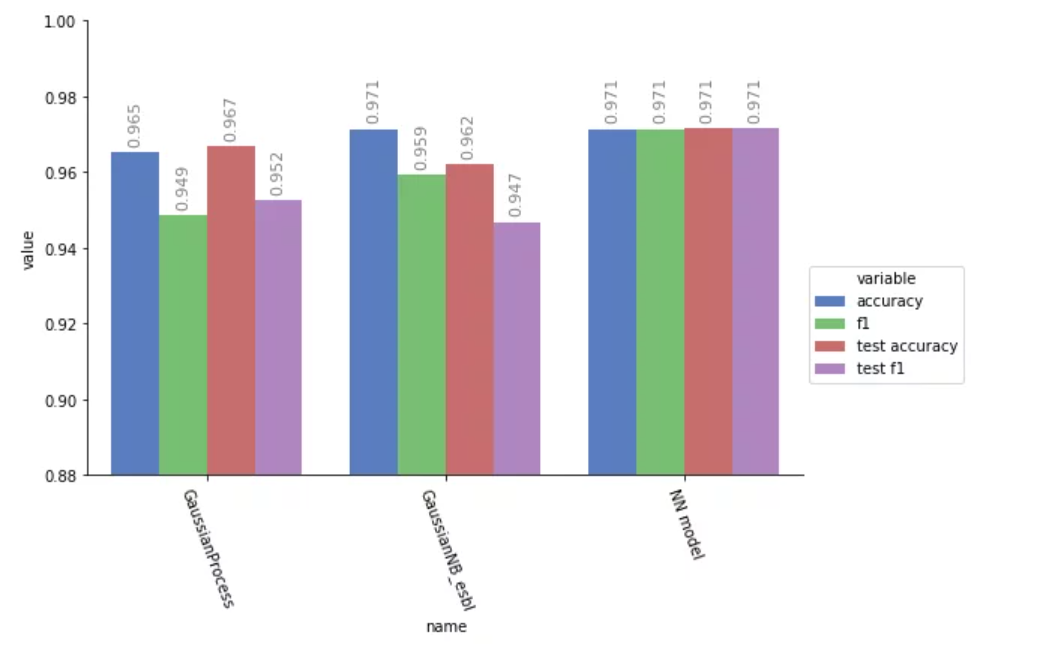
*Neural Network model architecture.*

This neural network classifier has resulted in 0.97 F1 mean scores on cross-validation. This new neural network model’s F1 score is better compared to the best model’s score gained in Step 3. Here are the top three models results so far.



*Cross-validation results of the top three models.*

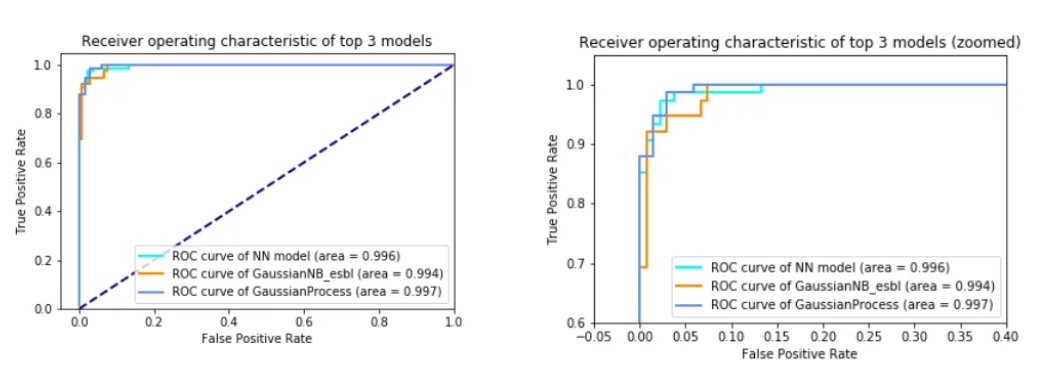
Now let’s evaluate these models on the test data set which previously was not shown to classifiers imitating new data. Below are the results demonstrating just how well these models performed on the test data set.



As it’s shown in the graph, neural network classifier have performed better by gaining 0.97 F1 scores on the test set.

**Step 5: Evaluating Output Quality Through Receiver Operating Curves**

In order to further evaluate classifiers’ output quality, let’s view their receiver operating characteristic (ROC) curves.



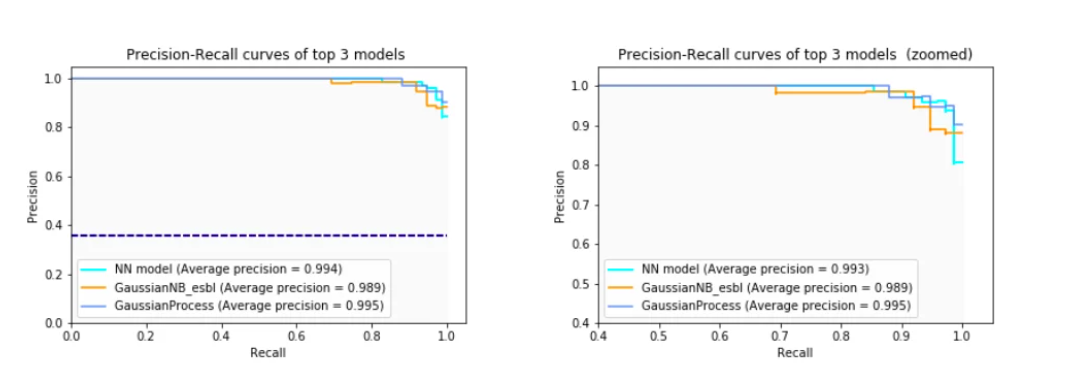
The ROC graph efficiency is measured by the area under the curve. An area of 1 represents a perfect classifier, an area of 0.5 represents a worthless classifier (navy color, dashed line in the graph). Here is the academic point system for judging classifiers efficiency given to area under the curve.

0.90-1 = excellent (A)  
0.80-0.90 = good (B)  
0.70-0.80 = fair (C)  
0.60-0.70 = poor (D)  
0.50-0.60 = fail (F)

As it’s shown in the graph, all of three classifiers have above 0.99 area under the curve which is considered excellent.

**Step 6: Evaluating Output Quality Through Precision-Recall Curves**

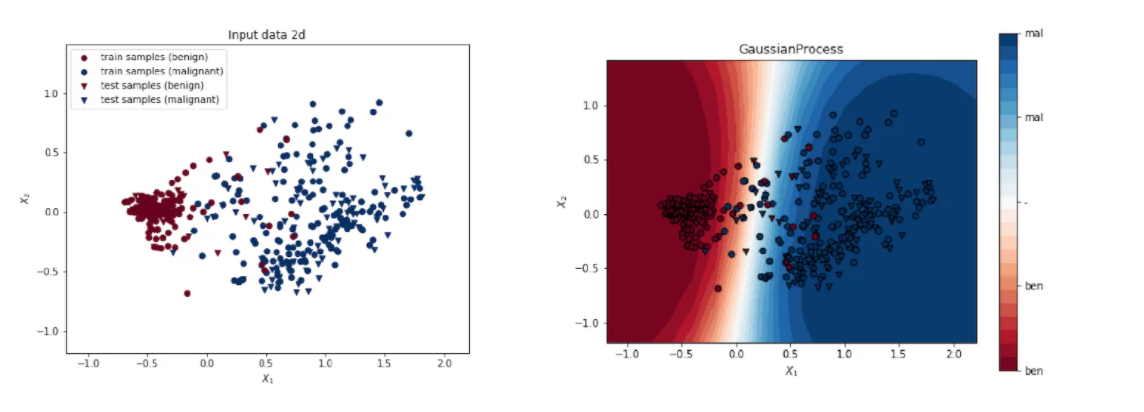
Let’s also look at the precision-recall curves associated with these classifiers.

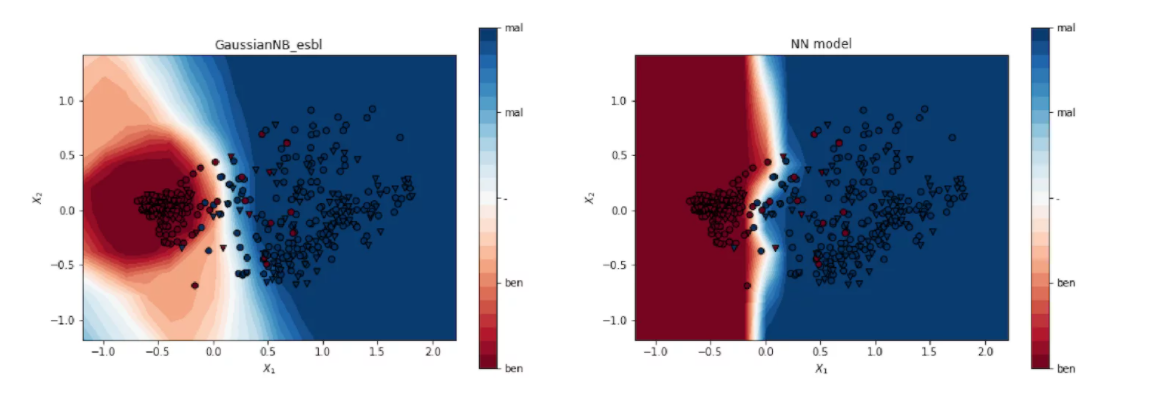


The navy dashed line represents the baseline, where the perfect model is the one with 1 average precision. As you can see, all three models’ average precision is close to 1, which are excellent scores.

**Step 7: Visualizing the Decision Boundaries**

Lastly, an additional note about the models’ decision boundaries:  
In order to gain some visual intuition about the data set and the algorithms decision boundaries, we will reduce the dimensionality of 9D feature space to 2D using PCA techniques and visualize the decision boundaries.





**Conclusion**

These models have shown excellent results on Breast Cancer Wisconsin (Diagnostic) Data Set, however, in order to trust the models, we need to further test them with new data and make sure they are still leading to excellent results. One possible weakness associated with these models is that they do not include any demography, race and genetic sequences attributes and other useful information that could potentially strengthen the ground for classification. One last note of caution: although the approaches outlined in this article may show promising results, the intention was to demonstrate the potential of AI algorithms, it was not intended for clinical use.

**References**

Citations for using [Breast Cancer Wisconsin (Diagnostic) Data Set](http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29)  
[1] O. L. Mangasarian and W. H. Wolberg: “Cancer diagnosis via linear programming”, SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.  
[2] William H. Wolberg and O.L. Mangasarian: “Multisurface method of pattern separation for medical diagnosis applied to breast cytology”, Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.  
[3]O. L. Mangasarian, R. Setiono, and W.H. Wolberg: “Pattern recognition via linear programming: Theory and application to medical diagnosis”, in: “Large-scale numerical optimization”, Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30.  
[4] K. P. Bennett & O. L. Mangasarian: “Robust linear programming discrimination of two linearly inseparable sets”, Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).