Application of SVM for breast cancer prediction based on Fine Needle Aspiration images features

Mr. A. Perdomo[a,c,*]

[a]VGTU, Dept. of Biomechanical Engineering, J. Basanavicius str. 28, LT-06287 Vilnius, Lithuania

ARTICLE INFO

Keywords:
Breast cancer
Classification
Support Vector Machines
Python
Scikit-Learn
Fine Needle Aspiration

ABSTRACT

Cancer is one of the main causes of death along the entire world and reports of the World Health Organization state that almost the 30-50% of the new diagnosed cases each year could be prevented. Thanks to Artificial Intelligence and machine learning, nowadays there are more support tools for doctors that can help to early detect and diagnose this type of disease, because here as in many other fields, times really matter. So, the purpose of this paper is explain the development of a first version of a support tool for early detect breast cancer, predicting the nature (malign or benign) of the studied breast mass.

1. Introduction

Cancer represents one of main causes of death along the entire world and despite of the huge advances that medical researchers are doing regarding treatment and cure, the number of cancer cases increases each year. The definition of cancer made by the World Health Organization [1] is the following: "Cancer is a large group of diseases that can start in almost any organ or tissue of the body when abnormal cells grow uncontrollably, go beyond their usual boundaries to invade adjoining parts of the body and/or spread to the organs." So, as the WHO states, the disease named as cancer can appear in almost any organ or tissue, but thanks to many diagnosis tools, some of them are easily diagnosed.

Due to the anatomical differences between males and females, there are certain cancer types that can only appear in each individual. For example, a female cannot suffer from prostate cancer. In the case of women, according to the 2018 Europe Globcan (Global Cancer Observatory) statistics [2], the breast cancer represented the 12.4% of the total number of new cancer cases diagnosed during 2018, affecting 522513 women in Europe, which represents the most common cancer diagnosed in the continent in the year 2018. Reducing the scope of the statistics, breast cancer was obviously the most common cancer diagnosis in women during the 2018, supposing the 26.4% of the total cancer diagnosis in women of all ages.

Despite breast cancer is not the most mortal type of cancer that affect humans, is the third just behind lung cancer and colon cancer [2]. But it should be noted that the breast cancer is positioned in the third place by mortality ratio of all types of cancer and it only affects women, so it is remarkable that a type of cancer that only affects approximately the 55% of the population due to its biological characteristics is positioned in the third place of that list [3].

Then, as it happens with all types of cancer, an early diagnosis is a key factor in the effect of the disease in the human being [4]. In this area there are two main tools that are widely use diagnose breast cancer:

- Mammography, which is a process that uses low energy x-rays to obtain an image of the inner breast and it is helpful to detect tumours, masses and different pathologies that can affect the human breast. The main issue is the high rate of false positives and false negatives in women with dense breasts [4].

- Fine Needle Aspiration (FNA), which is a procedure used to analyse the tissue of the tumour or mass and state its nature (malign or benign). The FNA uses a thin 25 gauge hollow needle [5] which is inserted in the mass which is going to be sampled, in this case breast masses. Then, after getting the sample from the mass tissue, the sample is

*Corresponding author
alberto.perdomo-garica@stud.vgtu.lt (A. Perdomo)
ORCID(s): xxxx-xxxx-xxxx-xxxx (A. Perdomo)

stained (which consist in wax the tissue to later fillet it) to be ready to be analysed under microscope. The complete procedure is usually known as FNAB, which stands for Fine Needle Aspiration Biopsy [6]. When doing the biopsy [7], the laboratory researcher analyses the cells, cells nuclei, etc… and its geometry, to recover numerical data which will be interpreted by a specialist in the field.

At this point is when information systems get into the track: Valuable experience is a grade that only specialists with thousands of hours developing a certain labour can have, and it is well known that in fields like medicine that experience can help the specialist to make a more accurate diagnosis. One of the approaches of information systems in medicine is not replace the doctor decision power but support it with intelligent systems to be able to make more precise and early diagnosis, through artificial intelligence techniques and algorithms.

Artificial Intelligence is defined as the computer or system simulation of intelligent human behaviour thanks to algorithms that can emulate the functioning of human brain, thinking, learning… And artificial intelligence can be divided in two main groups:

- Supervised methods: Develop models based on input and output data.

- Unsupervised methods: Also known as clustering technique, it develops models based only in input data.

Then, the purpose of this research is to develop a machine learning model, using a certain type of algorithm (SVM), which can help an oncologist to make an early diagnosis of breast cancer based on some features of the FNA images.

Support Vector Machine is highly convenient when working in high dimensional spaces and it is proved that it reports a great effectiveness, efficienty and accuracy. In this case, it was selected due to different factors: SVM presents the possibility to use a custom kernel, which is an algorithm which recognizes patterns in the data and can make the model more suitable for the current dataset. Also, the SVM is really recommended for two class datasets, which is the case of the data in this project, having only benign or malign cancer; and this is not as common as it may seen at a first sight, because usually medical tools contemplate the possibility of having a healthy patient, so then there is no binary classification because there are three classes.

Then, as the model was developed in Python, it contains a grid search based on cross validation (which will be explained in further chapters), to select the best set of parameters for the SVM model to reach the highest performance. So, this is very convenient too.

Finally, to sum up with the algorithm selection, SVM algorithm presents for this particular project and data, a great compromise between time, customization and computational cost, which makes it the best option. There are cases in which the selection of the model can take some time and tests due to the expected performance of the final solution, but in this case a general-purpose algorithm that could be adjusted is more than enough.

To conclude this section and as it is going to be explained in the next section, there is going to be used a pre-processed dataset, but for this project it has also been developed an algorithm which digitized an FNA image and measure the cell size using edge detection.

2. Methods

The digitization of an FNA image can be developed using different types of algorithms, and in this case it is going to be used a script which uses edge detection. It should be noted that due to the lack of a FNA images database, this algorithm will show how the procedure will be done using a large database of FNA images, but for training and testing purposes, there is going to be a pre-processed algorithm.

The developed script use different Python libraries that will be mentioned in later section and, using a reference cell in the image, it return the size of each of the cells present in the image. So there is a need of know at least one cell size in order to be able to get the other ones, but this can be solve placing a reference pattern which known size in the

moment of capturing the image from the biopsy microscope.

The dataset that is going to be used in this project has been downloaded from Kaggle [8]. It is a breast cancer dataset of Wisconsin city from 2016. This dataset is composed by features extracted from a digitized FNA image of a breast mass, so it has a description of the cell nuclei present in the images. The dataset itself is composed by the following features:

1. ID number: Patient number which connects the data with her electronic health record.
2. Diagnosis: Stating M for malignant and B for benign.

The next ten parameters are the computed real values for each cell nuclei (3-12):

a) Radius: Mean of distances from center to points on the perimeter.

b) Texture: Standard deviation of gray-scale values.

c) Perimeter.

d) Area.

e) Smoothness: Local variation in radius lengths.

f) Compactness: $\text{Perimeter}^2/area - 1.0$

g) Concavity: Severity of concave portions of the contour.

h) Concave points: Number of concave portions of the contour.

i) Symmetry

j) Fractal dimension: Coastline approximation – 1

The mean, standard error and worst/largest value (mean of the three largest values) were calculated for each image, resulting in 30 additional features. For instance, if the field 3 is the mean radius, field 13 is radius SE and field 23 is worst radius. The dataset is composed by 357 benign cases and 212 malign ones.

What a medical researcher sees when studying an FNA image can be seen in the figure 1, what in this case represents malign cancer cells [9].

In order to create a model using the described data that will predict the nature of a breast mass based on the explained features, the scripting language that is going to be used is Python [10]. Python is an open-source interpreted, high-level scripting language which is widely used in the scientific and development community due to the great number of possibilities and interoperability that it presents.

To be more precise, among other general-purpose libraries, the AI library that is going to be used in this project is called Scikit-Learn [11] and it provides open-source simple and efficient tools for predictive data analysis. In the following section, the Scikit-learn used SVM algorithm and the pseudo-code of the project will be presented.

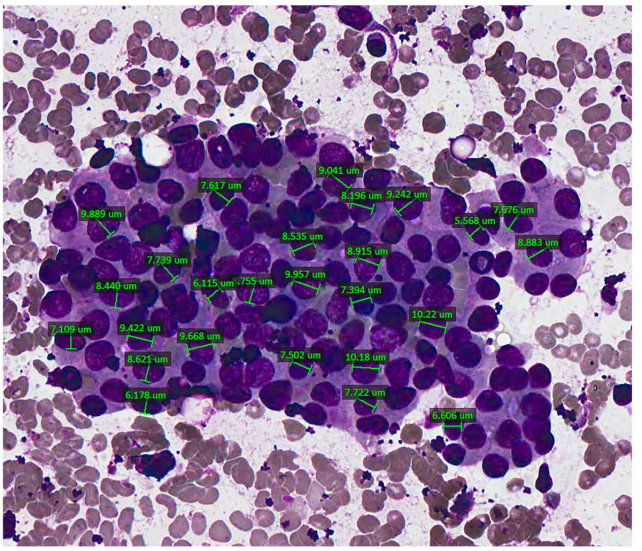

Figure 1: Biopsy of a FNA sample.

3. Theory and calculation

Earlier in this paper Artificial Intelligence was defined as the computer or system simulation of intelligent human behaviour thanks to algorithms that can emulate the functioning of human brain, thinking, learning, among others. So, there is no doubt about the huge amount of possible developed algorithms that can be used to emulate different human skills, since it is not the same think than learn, or feel than read.

The most important essential for training an Artificial Intelligence model is the data. Usually, when more quality data is feed to the algorithm, more accurate model is going to be developed. As it was stated in previous sections, the model in this case is going to use a pre-processed dataset, but as it was also advanced, this project includes the development of an algorithm based on edge detection to demonstrate how the data would be extracted from an FNA image.

In the pseudo-code in figure 2 can be appreciated that the size is print in the terminal and in the image; that is only for testing purposes. In the real application of this algorithm, the data should fill a csv file or a Pandas data frame. Further evaluations of the algorithm's results will be presented in following chapters.

In the prediction field, Artificial Intelligence can be used to predict different parameters such as words, colours, numbers, classes... but all of them have in common the procedure of modelling which appears in figure 3. Since the given dataset has already labels [8], as it was explained, the developed model in this project is going to predict labels, and since there are only two labels, the prediction algorithm can be generally named a binary classifier, because it will classify each sample composed by 30 features in one of the two labels (0 or 1).

The output data will be part of the model so the chosen algorithm will be a supervised one. For that algorithm, and explaining more the figure 4, the training part consist of the learning part of the algorithm, in which it will analyse the features and check the expected output (label) to learn and find relations between the data, in order to create a model. The testing part consist in checking how the model works for unseen data: The model will be feed with data that it has never seen and it will generate the output (so in this case the labels are not part of the input of the algorithm); then with the known correct labels and the model output, the accuracy of the algorithm can be calculated.

```
# imports

# load the image and convert it to grayscale
fna_image = cv2.imread(args["fna_image"])
gray = cv2.cvtColor(fna_image, cv2.COLOR_BGR2GRAY)

# perform edge detection
# find contours in the edge map
cnts = cv2.findContours(edged.copy(), cv2.RETR_EXTERNAL,cv2.CHAIN_APPROX_SIMPLE)
# sort them from left-to-right

# initialize the pixels per metric calibration variable
# loop over each contour
for c in cnts:
    # if the contour is not sufficiently large, ignore it

    # compute the rotated bounding box of the contour

    # order the points in the contour such that they appear
    # in top-left, top-right, bottom-right, and bottom-left order
    box = perspective.order_points(box)
    cv2.drawContours(orig, [box.astype("int")], -1, (0, 255, 0), 2)

    # using the ordered points of bounding box, compute the midpoint
    # between the top-left and top-right coordinates, followed by
    # the midpoint between bottom-left and bottom-right coordinates

    # draw lines between the midpoints

    # compute the euclidean distance between the midpoints and object size

    # draw the object sizes on the image and print the size
```

Figure 2: Edge detection algorithm pseudocode.

Accuracy is one of the metrics for evaluating classification methods, and formally it has the following definition:

$$Accuracy = \frac{Numberof\,correctpredictions}{Totalnumberof\,predictions}$$

For binary classification, accuracy can be calculated in terms of positives and negatives as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP is True Positive and states that the condition exists when it really exists; TN is True Negative and states that the condition does not exist when it does not; FP is False Positive and states that the condition exists when it does not; and FN is False Negative and states that the condition does not exist when it really exist.

It should be noted that this is not the only metric that is used to measure the performance of a classification algorithm, but it is the main one.

In order to achieve the set goal, the algorithm that is going to be used is known as Support Vector Machines or Support Vector Networks [11]. This algorithm is composed by a set of supervised learning methods used for classification, regression and outliers detection.

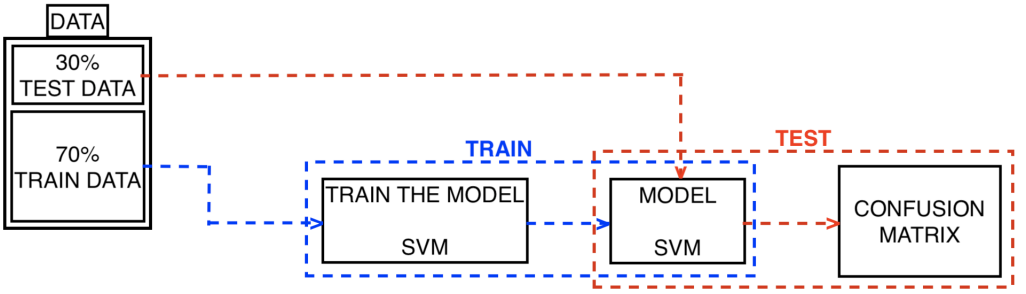

Figure 3: Typical train-test split for machine learning models.

SVM are highly convenient when working in high dimensional spaces and it reports a great effectiveness. The name of the algorithm comes from the use support vector, which are a subset of training points in the decision function, so this algorithm is memory efficient. On the contrary, if the number of features is much greater than the number of samples, this algorithm needs a more precisely set up with custom kernel functions (the function used to predict a score to select the label), which can be pretty difficult in order to obtain a well performing model.

In this project, apart for creating the model, it has been used a method called Grid Search in order to find the best estimator. An estimator is a function which "picks up" the best model, and it is based in a serie of parameters. It relies its performance in the other method called cross validation [12], which divides the training split into some sub splits and use each of them to test different models and then uses the estimator to choose the best ones.

The pseudo-code of the developed model is shown in the following figure 4.

,

```
# imports
# import the data
cancer = pd.read_csv('breast_cancer_data.csv')

# split the data in X and y
# encode class values as integers (0==B or 1==M)
# split the data in training and testing
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)

# create and train the model
svm = SVC()
svm.fit(X_train, y_train)
predictions = svm.predict(X_test)

# print the first results for the model with standard parameters

# grid search for finding the best parameters for the model
param_grid = {'C':[0.1, 1, 10, 100, 1000], 'gamma':[1, 0.1, 0.001, 0.0001]}

# make the predictions with the best parameters
grid_predictions = grid.predict(X_test)

# print the results for the best model with custom parameters
# plot non-normalized confusion matrix
```

Figure 4: SVM model pseudocode in Python.

4. Results

In previous chapters it has been stated that the expected results should be a predicted label, in order to give a diagnosis to the set of features of each test patient. It should be noted that the results presented in this section are guiding results but not decision making ones. There are features that a mathematical model cannot have as an input but the medical researcher can consider, so the developed model and the results obtained from it are a support tool for the doctor, but the decision of the nature of the breast mass will always remain in the power of the doctor.

Before continuing further in this section, two more performance metrics need to be described since they are going to be referenced in the following paragraphs [13]:

- Precision: Expresses the portion of datapoints that the model predicts as relevant which actually were relevant, for each class.
- Recall: Expresses the ability of the model to find all the relevant instances in the dataset, for each class.

The results section is composed by: The results of the cell measuring algorithm based on edge detection, the very first results of the model with standard parameters and the results for the best set of parameters found by grid search.

Regarding the results of the cell measuring algorithm, it would be an error to talk from the quantitative point of view since there are no prior data about the cell size of the cells in the image, so the evaluation is going to be mainly quantitative. In the following figure number 5, it can be observed that the algorithm has two results, the desirable one in the left and the unwanted one in the right. The discussion of these results will be done in the following section.

,

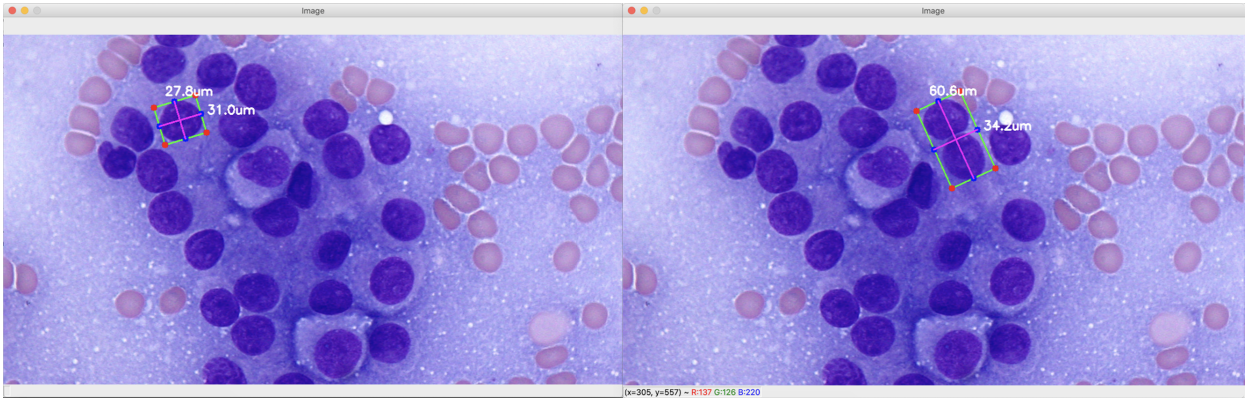

Figure 5: Cell size measured using the edge detection algorithm.

The first model with standard parameters gave the following results which appear in table 1, and the best model results with custom parameters gave the following appear in table 2.

	First results	
Metric		Score
Recall	0	1.0
	1	0.79
Precision	0	0.89
	1	1.0
Accuracy		0.9238

Table 1
First results of the model with standard parameters.

	Best results	
Metric		Score
Recall	0	0.97
	1	0.89
Precision	0	0.49
	1	0.95
Accuracy		0.9415

Table 2
Best results of the model with custom parameters.

For the best model, the confusion matrices are the following in figure 6. It should be noted the confusion matrix on the left is the non-normalized confusion matrix, which means that it represents the number of predictions; however, the matrix in the right side is normalized and it represents the percentage of each prediction.

,

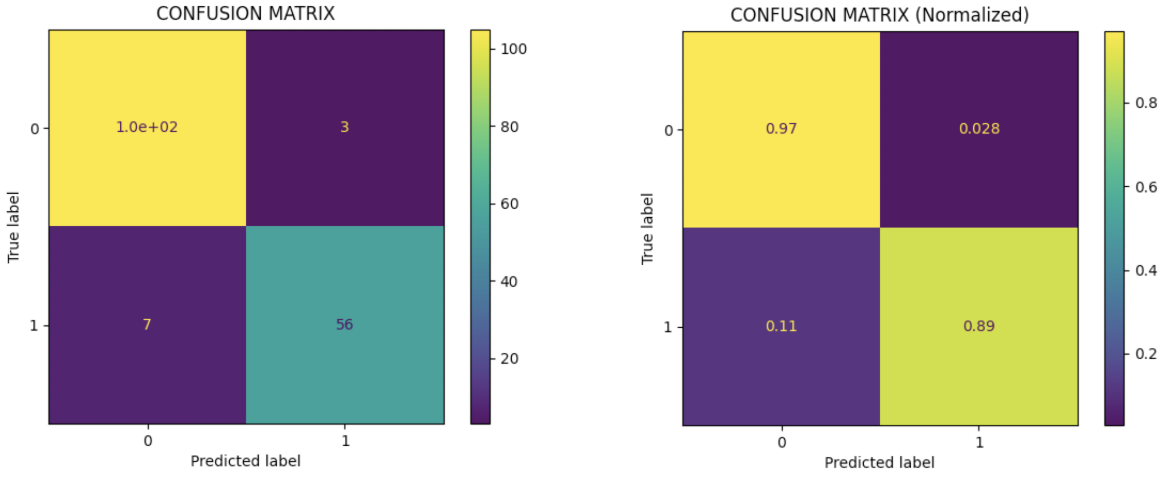

Figure 6: Confusion matrices for the best model obtained with custom parameters.

5. Discussion

In this section, the results presented in the last section are going to be interpreted. Before continuing with deeper evaluation of the results, it should be noted that a model trained with 569 patients gives representative results but in order to consider this model as a support tool for doctors, it will need to be trained with different extreme cases, with much more data and with finer parameters.

Regarding the script for measuring cell size, it should be stated that the results were not what it can be evaluated as good, due to different factors: First of all, there was not found a database of breast cancer FNA images in order to check the algorithm performance in other images. Secondly, when two edges of two objects are relatively closed, the algorithm does not differentiate the two objects, so the measured "cell" are in fact two or more cells. Thirdly, the working procedure of this algorithm must be improved in order to measure different sizes along the cell, because the cells are not perfectly circular so it should measure different diameters for the same cell to be able to get the worst (larger) diameter, among other parameters.
However, for the purpose of this project and the time used to develop this algorithm, the results are very promising and then, it is open a new research line to continue improving this algorithm. The purpose of that script is to show how the parameters extraction should be done.

Regarding the predicting model, it needs to be differentiated what is considered as a good performance and what is scientifically accepted as a good performance. Thanks to the grid search, the best model was found and reported an accuracy of 0.9415 against the 0.9239 of the standard model. Although there is no doubt that this is a high accuracy for such a simple model, it needs to be clarified that these results are far from what would be suitable for a medical support tool.

Paying attention to those metrics, the model can predict the correct label in 94 women out of 100, which seems to be a high ratio. But if we consider that in the case of predicting the class of malign mass (or what is the same, having malign cancer) there are 6 women that would be told not having it when they really do. So, when the target class affect people the error rate should be close to zero, not a 6%.

In conclusion, it needs to be mentioned that the general performance of the predicting model is what we would consider good, having a high accuracy, but considering that the model predict class that can derive in death of human beings, that accuracy is not high enough.

Finally, it is open a future research line in which the model should be fitted better and with finer parameters (even using another machine learning algorithm) in order to be able to be consider it as a support tool for clinical diagnosis.

References

[1] WHO, "cancer @ www.who.Int," 2018.

[2] International Agency for Research on Cancer, "Europe Statistics 2018," *The Global Cancer Observatory*, vol. Population, pp. 1–2, 2019.

[3] Statisca, "european population @ www.statista.com."

[4] L. Wang, "Early diagnosis of breast cancer," *Sensors (Switzerland)*, vol. 17, no. 7, 2017.

[5] N. Mateša, N. Dabelić, I. Tabain, and Z. Kusić, "Fine needle aspiration of the thyroid," *Acta Clinica Croatica*, vol. 41, no. 2, pp. 123–131, 2002.

[6] T. Mu and A. K. Nandi, "Breast cancer detection from FNA using SVM with different parameter tuning systems and SOM-RBF classifier," *Journal of the Franklin Institute*, vol. 344, no. 3-4, pp. 285–311, 2007.

[7] K. Yamashiro, N. Yoshimi, T. Itoh, H. Takino, M. Nakajima, M. Azuma, K. Taira, S. Makio, S. ichi Shiina, S. Hata, S. Urabe, J. Fukuoka, and I. Mori, "A small-scale experimental study of breast FNA consultation on the internet using Panoptiq," *Journal of the American Society of Cytopathology*, vol. 8, no. 4, pp. 175–181, 2019.

[8] Kaggle, "breast-cancer-wisconsin-data @ www.kaggle.com."

[9] American Cancer Society, "acsjournals.onlinelibrary.wiley.com."

[10] Python, "www.python.org," 2015.

[11] SciKit Learn, "plot_grid_search_digits @ scikit-learn.org."

[12] Amazon Web Services, "cross-validation @ docs.aws.amazon.com."

[13] Towardsdatascience.com, "towardsdatascience.com."