

High Accuracy Classification of Populations with Breast Cancer: SVM Approach

Philip de Melo¹, Mane Davtyan²

¹Computer Science Department, Bowie State University, Bowie, USA

²College of Science and Engineering, American University of Armenia, Yerevan, Armenia

Email address:

pdemelo@bowiestate.edu (Philip de Melo), mane_davtyan@edu.aua.am (Mane Davtyan)

To cite this article:

Philip de Melo, Mane Davtyan. High Accuracy Classification of Populations with Breast Cancer: SVM Approach. *Cancer Research Journal*. Vol. 11, No. 3, 2023, pp. 94-104. doi: 10.11648/j.crj.20231103.13

Received: July 1, 2023; **Accepted:** July 27, 2023; **Published:** August 15, 2023

Abstract: Breast cancer is one of the most common cancers diagnosed in the United States. Breast cancer can occur in both men and women. The number of deaths associated with this disease is steadily declining, largely due to factors such as earlier detection and a new personalized approach to treatment. In this article, we offer a highly accurate and reliable classification approach based on feature engineering and an improved support vector machine (SVM) classifier. We examine a dataset with 30 features and use in-depth data analytics and visualization to pinpoint the top nine features that have a significant impact on classification accuracy. The SVM classification outperformed other classifiers, including kernel extensions, with a high accuracy of 99.12%. The study stresses the value of machine learning in medical diagnosis, notably in the early detection of breast cancer, and indicates the possibility for further research in this area utilizing deep learning architectures. Early detection of breast cancer is critical, and our findings contribute to the growing body of knowledge in this area, opening new avenues for improving cancer diagnosis and patient care.

Keywords: Breast Cancer, Support Vector Machine, Feature Engineering, Early Detection, Machine Learning, Classification, Data Analytics

1. Background

Breast cancer is the most common cancer affecting women and is the most diagnosed cancer worldwide. Breast cancer is also the most common cancer in women in the United States, except for skin cancers. It is about 30% (or 1 in 3) of all new female cancers each year.

The American Cancer Society's estimates for breast cancer in the United States for 2023 are as follow:

1. About 297,790 new cases of invasive breast cancer will be diagnosed in women.
2. About 55,720 new cases of ductal carcinoma in situ (DCIS) will be diagnosed.
3. About 43,700 women will die from breast cancer.

Breast cancer mainly occurs in middle-aged and older women. The median age at the time of breast cancer diagnosis is 62. This means that half of the women who developed breast cancer are 62 years of age or younger when they are diagnosed. An exceedingly small number of women

diagnosed with breast cancer are younger than 40-45.

One of the most important in breast cancer treatment is its timely detection. Early-stage cancer detection could significantly reduce breast cancer death rates. The most critical point for the best prognosis is to identify early-stage cancer cells.

Computer science and machine learning, in particular, has emerged as a valuable tool to detect various medical conditions with greater accuracy than other approaches. Machine learning involves the creation of algorithms to classify patients with cancer or cancer-free.

Why advanced algorithms have become the forefront of cancer research? Screening for breast cancer is a very sensitive matter: aggressive screening strategies will maximize the benefits of early detection, whereas less-frequent screenings will reduce false positives, anxiety, and costs for those who will never even develop breast cancer. Using the technology presented in this and other papers may significantly slash the number of useless screenings thus saving considerable amount of money for this nation's healthcare.

Besides, current clinical guidelines use risk models to determine which patients should be recommended for supplemental imaging and MRI. Some guidelines use risk models with just age to determine if, and how often, a woman should get screened; others combine multiple factors related to age, hormones, genetics, and breast density to determine further testing. Despite decades of effort, the accuracy of risk models used in clinical practice remains largely inaccurate.

We propose a novel highly accurate, and robust classification algorithm based on an optimized support vector machine classifier and a random forest/feature engineering.

2. Introduction

Cancer is an uncontrolled growth of cells in the body that can rapidly spread to any organ and 90% of cancer patients die from metastasis. Numerous types of cancer exist, but lung cancer, breast cancer (BC), and skin cancer are the most prevalent. According to World Health Organization (WHO) reports, the cancer death ratio is as high as 9.2 million for lung cancer, 1.7 million for skin cancer and 627,000 for breast cancer.

Breast cancer is considered a multifactorial disease and the most common cancer in women worldwide (about 1.5 million women are diagnosed with breast cancer each year, and on average 500,000 women die from this disease in the world). Over the past 30 years, this disease has increased, while the death rate has decreased due to mammography screening.

Several image-guided deep-learning models were developed for the prediction of cancer. Along these lines, several machine-learning algorithms were utilized to distinguish benign from malignant cells based on histopathological reports (classification problems). Histopathological reports are stored in EHR from the time of diagnosis until the time of discharge, employing a text mining framework to extract meaningful information from medical records or text documents and apply it to machine-learning algorithms for cancer prediction.

Diagnostic mammography can assess abnormal breast cancer tissue in patients with subtle malignancy signs. Due to a large number of images, this method cannot effectively be used in assessing cancer-suspected areas. Approximately 50% of breast cancers were not detected in screenings while a quarter of women with breast cancer are diagnosed negatively.

Based on data, a stage-specific interpretation system was designed and this information serves as the primary resource for guiding patients' treatment methods. Following confirmation of the disease's stage and subtype, the healthcare provider initiates chemotherapy to mitigate the growth of cancer cells. This can be done by modifying the expression of several genes. Text mining has helped to find biologically relevant alternative therapeutic candidates. It is also true that drug development remains a lengthy and expensive procedure.

Most mammography-based breast cancer screenings are performed at regular intervals - usually annually or every two

years. On the other hand, experts suggest that considering other risk factors along with mammography screening can help in a more accurate diagnosis of women at risk. Moreover, effective risk prediction through modeling can not only help radiologists in setting up personal screening for patients and encouraging them to participate in the program for early detection but also help identify high-risk patients.

Machine learning and data mining represent modeling approaches, discovering hidden relationships to predict different diseases. A major challenge in predicting breast cancer is the creation of a model for addressing all known risk factors that influence the disease progression. Unfortunately, current prediction models only focus on the analysis of mammographic images without other critical factors such as lifestyle or laboratory data and patient biopsy.

Combining multiple risk factors in modeling breast cancer prediction could help the early diagnosis of the disease with necessary care plans. Collection, storage, and management of different data and intelligent systems based on multiple factors for predicting breast cancer are effective in disease management.

Therefore, multifactorial models with many risk features can be effective in assessing the risk of breast cancer. The current study aimed to predict breast cancer using different machine-learning approaches considering various factors in modeling.

Support Vector Machine is introduced by Vapnik [1-3]. SVM is a supervised learning technique used for regression and classification. the goal of SVM is to identify the specific hyperplane with the maximum margin that may divide the classes in a linear fashion. Finding data sets when there are insufficient training data and where the optimal solution cannot be guaranteed by the regular application of a large number of statistics is the aim of supporting vector machine learning.

Osareh and Shadgar considered support vector machines, K-nearest neighbours and probabilistic neural networks classifiers are combined with signal-to-noise ratio feature ranking, sequential selection-based feature selection and principal component analysis feature extraction to distinguish between the benign and malignant tumors of breast [4]. The best overall accuracy for breast cancer diagnosis was achieved equal to 98.80% and 96.33% respectively using support vector machines classifier models against two widely used breast cancer benchmark datasets.

Fan J, et al introduced a new method for breast cancer risk prediction. Oyewola D., et al developed a method for the prediction of biopsy results. Hou et al., considered an ML technology for assessing breast cancer in China. Behravan H. et al used ML and genetic and demographic factors for breast cancer prediction. [5-8]

Kumar K. et al, Alghunaim S, Al-Baity H. H. used ML for breast cancer prediction in various populations and in big data context. Asri H., et al, Lotfnezhad Afshar H, and Tapak L., et al., developed methods for breast cancer prediction and survival ratio. [9-13]

The breast cancer datasets are used in this study and may be found at the ACM SIGKDD Cup 2008 and the UCI

machine learning repository, respectively the former is a dataset of relatively small size, consisting of 577 data samples, each of which includes 32 distinct features.

3. Support Vector Machine: Original and Dual Formulation

Support Vector Machine (SVM) basically helps classify the data into two categories (cancer-yes and cancer-no) with the help of a multi-dimensional boundary to differentiate the outcomes. Let us consider a linear separator that can be expressed as:

$$W^T * \Phi + b = 0 \quad (1)$$

where W defines the slope of a hyperplane and b is the intercept (bias), and Φ is a coordinate vector. In 2-D: $\Phi = \text{col}(x, y)$ and $W = (m, -1)$

$$mx - y + b = 0 \quad (2)$$

which is the equation of line in a 2-D Euclidean plane.

$$W^T * \Phi + b > 0 \quad (3)$$

Support Vectors represent the points that closest to the hyperplane. A separating line will be defined with the help of these data points. The distance between the hyperplane and the support vectors is called margin. For all points below the decision hyperplane the following is true:

$$W^T * \Phi + b < 0 \quad (4)$$

The distance from a point in the Euclidean space to the hyperplane can be presented by the following formula:

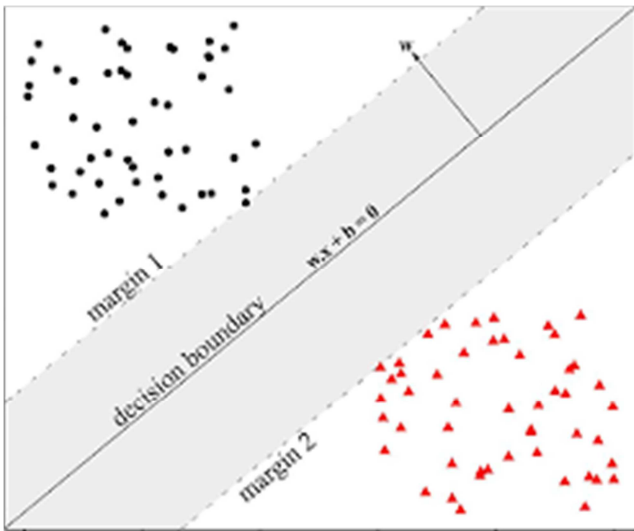


Figure 1. Decision boundary.

Positive hyperplane:

$y - mx - b = 1$

Negative hyperplane

$y - mx - b = -1$

Decision hyperplane

$y - mx - b = 0$

$$D = \frac{|W^T * \Phi + b|}{\|W\|} \quad (5)$$

where

$$\|W\| = \sqrt{\sum_i w_i^2} \quad (6)$$

Is a measure in L_2 norm. Let us consider now the optimization problem. We will consider the first case when we would like to maximize the distance between two hyperplanes:

$$W^T * \Phi + b = 1 \quad (7)$$

and

$$W^T * \Phi + b = -1 \quad (8)$$

The distance between 2 hyperplanes can be expressed as:

$$D_h = \frac{2}{\|W\|} \quad (9)$$

Maximizing the distance between to hyperplanes is equivalent to the minimization of the inverse expression:

$$D_{max} = \min_w \frac{\|W\|}{2} \quad (10)$$

Under the following constrains:

$W^T * \Phi + b > 1$ for all “positive” points

$W^T * \Phi + b < -1$ for all “negative” points

Let us introduce the classifier y that defines positive and negative points.

$$y = \begin{cases} 1 & \text{if } W^T * \Phi + b \geq 0 \\ -1 & \text{if } W^T * \Phi + b < 0 \end{cases} \quad (11)$$

So the optimization problem can be presented as finding solution of the following problem find optimal w^* and b^* by the minimization of the following expression:

$$\min Q_{b,w,\xi}(W, \xi) = \left(\frac{W^T W}{2} \right) + C \sum_i \xi_i, \quad (12)$$

This constraint: $y_i [W^T * \Phi_i + b] \geq 1 - \xi_i, i = 1 \dots N$ is very important as it requires that all the training points are correctly classified. This is the constrained problem that can be solved via the introduction of Lagrange multipliers. So the bottom line is that Lagrange multipliers is really just an algorithm that finds where the gradient of a function points in the same direction as the gradients of its constraints, while also satisfying those constraints. For soft bounds the cost function changes:

$$\min Q_{b,w,\xi}(W, \xi) = \left(\frac{W^T W}{2} \right) + C \sum_i \xi_i, \quad (13)$$

if $y_i [W^T * \Phi_i + b] \geq 1 - \xi_i$

The idea is: for every vector Φ_i , we introduce a variable ξ_i . Its value is the distance of Φ_i , from the corresponding class's margin if Φ_i , is on the wrong side of the margin, otherwise zero. Thus the points that are far away from the margin on the wrong side would get more penalty.

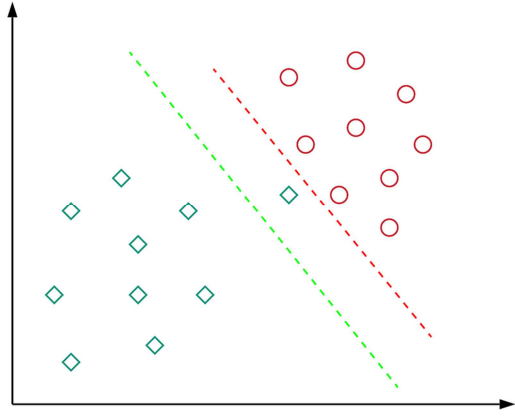


Figure 2. Soft margins allow misclassifications, however they reduce the overfitting.

4. Support Vector Machine: Fenchel Transform and Lagrange Multipliers

Suppose we have a cost function that is convex:

$$f(w) = \frac{\|w\|^2}{2} \text{ and a constraint}$$

$$g(\Phi, W) = y_i [W^T * \Phi_i + b] \geq 1 - \xi_i \quad (14)$$

We can see that there is only one point where two vectors point in the same direction: it is the minimum of the objective function, under the constraint. Here, the left-hand side of the inequality could be thought of like the confidence of classification. Confidence score ≥ 1 suggests that classifier has classified the point correctly. However, if confidence score < 1 , it means that classifier did not classify the point correctly and incurring a linear penalty of ξ_i . By definition, Lagrange multiplier is a λ parameter that relates the gradient of the cost function to the gradient of the constraint:

$$\nabla f(w, b) = \lambda \nabla g(w, b) \quad (15)$$

Let us form the Lagrangian:

$$L(w, b, \lambda) = \frac{\|w\|^2}{2} - \sum_i \lambda_i [y_i (W^T * \Phi_i + b)] + \xi_i - 1 \quad (16)$$

The Lagrangian for hard bound can be expressed by:

$$L(w, b, \lambda) = \frac{w^T * w}{2} - \sum_i \lambda_i [y_i (W^T * \Phi_i + b)] - 1 \quad (17)$$

with $\lambda_i \geq 0, i = 1, \dots, N$

Let us apply now ∇ to (18). This yields:

$$\nabla L(w, b, \lambda) = \nabla f(w, b) - \lambda \nabla g(w, b) \quad (18)$$

The cost function is maximizes when the gradient to it is colinear to the gradient to the constraint. This means that $\nabla L(w, b, \lambda) = 0$ which is equivalent to a set of equations:

$$\nabla L(w, b, \lambda) = 0 \rightarrow \begin{cases} \frac{\partial L}{\partial w} \\ \frac{\partial L}{\partial b} \\ \frac{\partial L}{\partial \lambda} \end{cases} = 0 \quad (19)$$

Let us explicitly calculate the Lagrangian derivatives:

$$\frac{\partial L}{\partial w} = W - \sum_i \lambda_i y_i \Phi_i = 0 \quad (20)$$

$$\frac{\partial L}{\partial b} = \sum_i \lambda_i y_i = 0 \quad (21)$$

Resulting classifier:

$$y(x) = \text{sign}[\sum_i \lambda_i y_i \Phi_i^T \Phi_i + b]$$

The Lagrangian in the dual space can be rewritten as:

$$\max_{\lambda} L(\lambda) = \sum_i \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j \Phi_i^T \Phi_j \quad (22)$$

We have to maximize the Lagrangian subject to

$$\lambda_i > 0, \text{ for all } i \text{ and } \sum_i \lambda_i y_i = 0 \quad (23)$$

$\lambda_i > 0$ Φ_i defines a support vector, if $\lambda_i = 0$, Φ_i is not a support vector.

To solve the actual problem we do not require the actual data point instead only the dot product between every pair of a vector.

To calculate the “b” biased constant we only require dot product.

The major advantage of dual form of SVM over Lagrange formulation is that it only depends on λ .

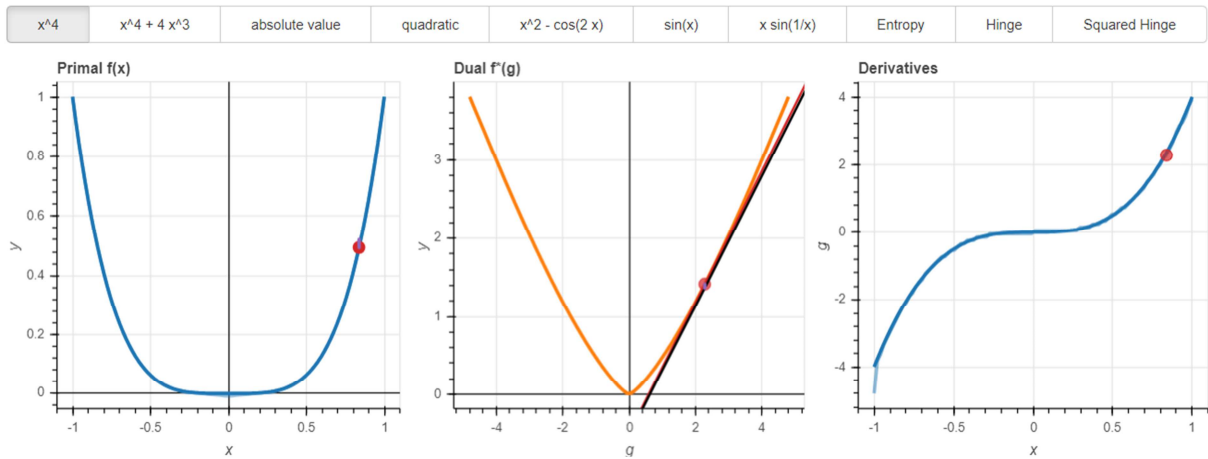


Figure 3. Fenchel dual transform transfers points on a curve on slopes of a convex conjugate function.

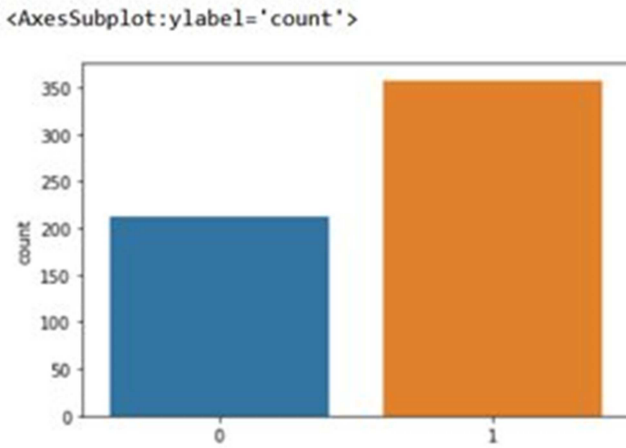


Figure 4. Diagram of benign and malignant instances.

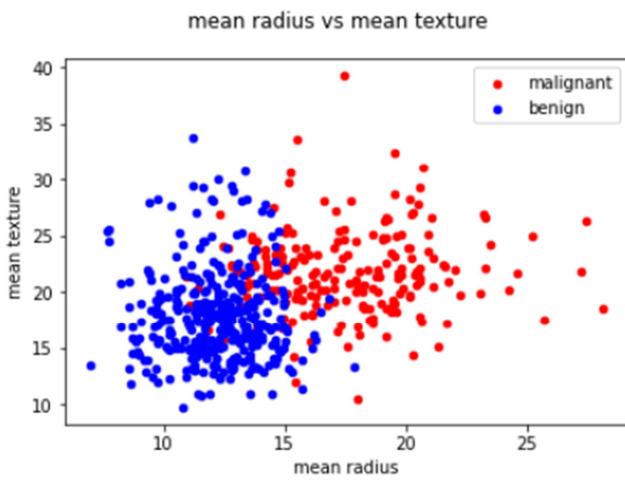


Figure 5. Mean radius vs. mean texture (red dots vs. blue dots-benign).

The Fenchel dual transfer results in a convex function.

The Lagrangian for the soft bounds in the dual form can be expressed as:

$$Q(W, b, \xi, \lambda, \mu) = \left(\frac{W^T W}{2} \right) + C \sum_i \xi_i - \sum_i \lambda_i [y_i (W^T * \Phi_i + b)] - 1 + \xi_i - \sum_i \mu_i \xi_i \quad (24)$$

This is a minimax problem

$$\max_{\lambda, \mu} \min_{W, b, \xi} Q(W, b, \xi, \lambda, \mu) \quad (25)$$

5. Description of Data

The data has 30 main features out of which 29 are numerical, while the only categorical column is diagnosis. Diagnosis column is a target column, in a sense of resulting into either M - Malignant or B - Benign, depending on other 29 numerical column values.

Number of Instances: 569

Number of Attributes: 30

Attribute Information:

radius (mean of distances from center to points on the perimeter)

texture (standard deviation of gray-scale values)

perimeter

area

smoothness (local variation in radius lengths)

compactness (perimeter² / 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")

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

class: WDBC-Malignant, - WDBC-Benign

Fundamental data statistics

Table 1. Summary for numerical features.

Feature	Min	Max
radius (mean)	6.981	28.11
texture (mean)	9.71	39.28
perimeter (mean)	43.79	188.5
area (mean)	143.5	2501.0
smoothness (mean)	0.053	0.163
compactness (mean)	0.019	0.345
concavity (mean)	0.0	0.427
concave points (mean)	0.0	0.201
symmetry (mean)	0.106	0.304
fractal dimension (mean)	0.05	0.097
radius (standard error)	0.112	2.873
texture (standard error)	0.36	4.885
perimeter (standard error)	0.757	21.98
area (standard error)	6.802	542.2
smoothness (standard error)	0.002	0.031
compactness (standard error)	0.002	0.135
concavity (standard error)	0.0	0.396
concave points (standard error)	0.0	0.053
symmetry (standard error)	0.008	0.079
fractal dimension (standard error)	0.001	0.03

Feature	Min	Max
radius (worst)	7.93	36.04
texture (worst)	12.02	49.54
perimeter (worst)	50.41	251.2
area (worst)	185.2	4254.0
smoothness (worst)	0.071	0.223
compactness (worst)	0.027	1.058
concavity (worst)	0.0	1.252
concave points (worst)	0.0	0.291
symmetry (worst)	0.156	0.664
fractal dimension (worst)	0.055	0.208

6. Data Analytics and Visualization

In this section, we will display a few characteristics of the data features.

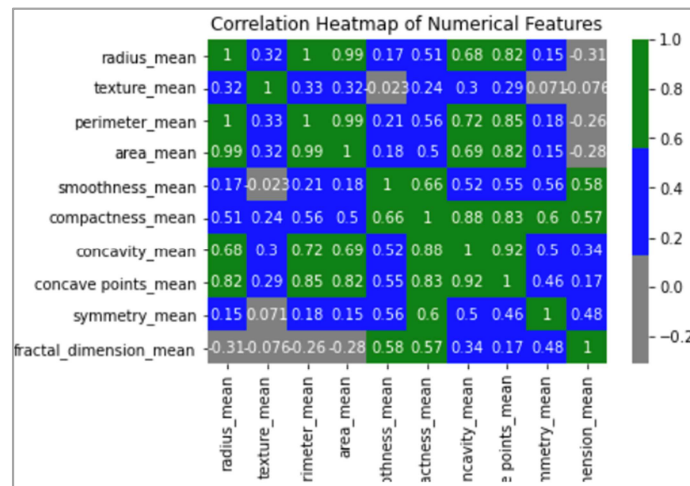


Figure 6. Heat map shows the correlation values between features.

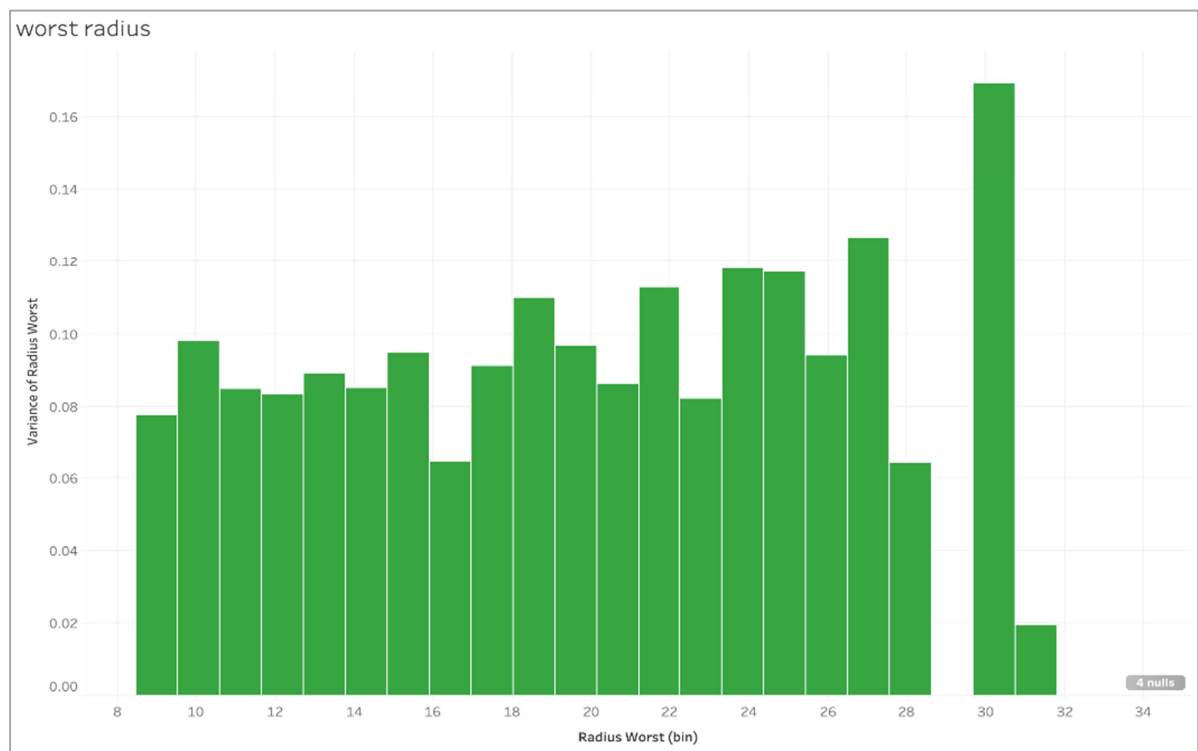


Figure 7. Variance of Concave points worst shows large deviations at its lowest values. The variance is mean squared difference between each data point and the center of the distribution measured by the mean. It plays an important role in the data science showing the credibility of data collection.

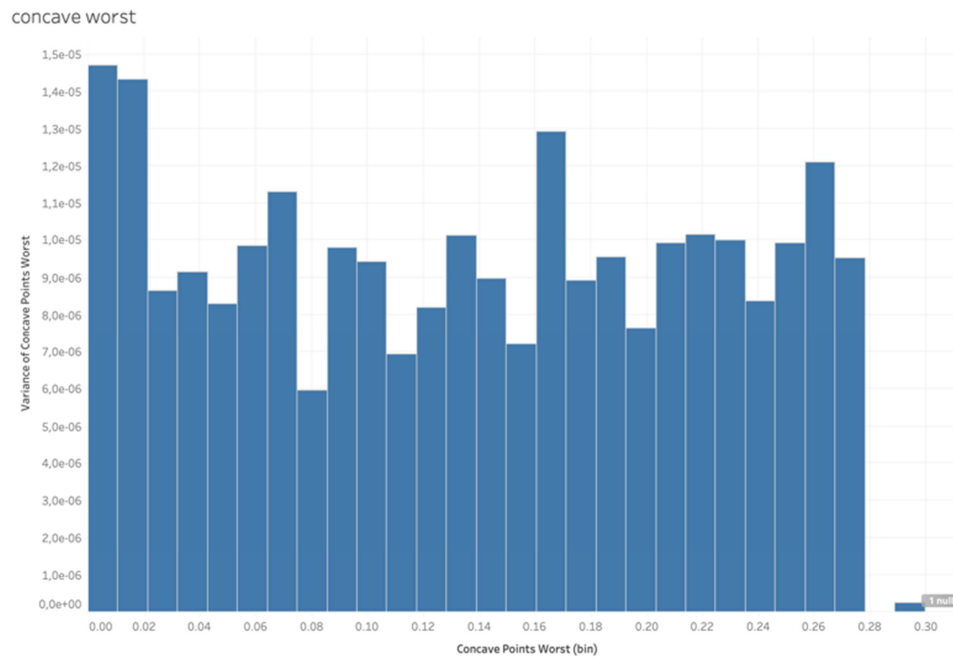


Figure 8. Variance of radius worst. It shows that variance in this column is the largest at around 30.

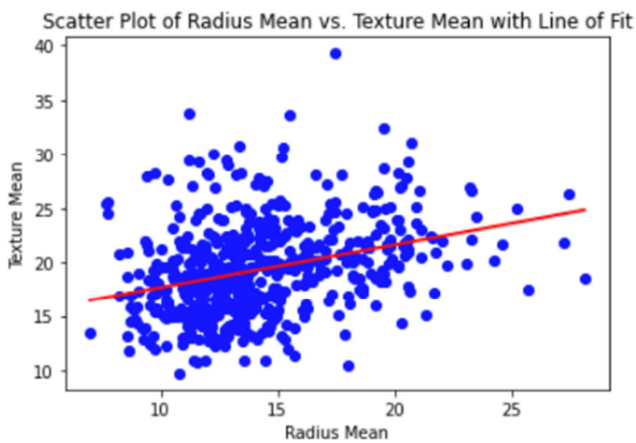


Figure 9. The scatter plot shows correlation between two features: radius mean and texture mean.

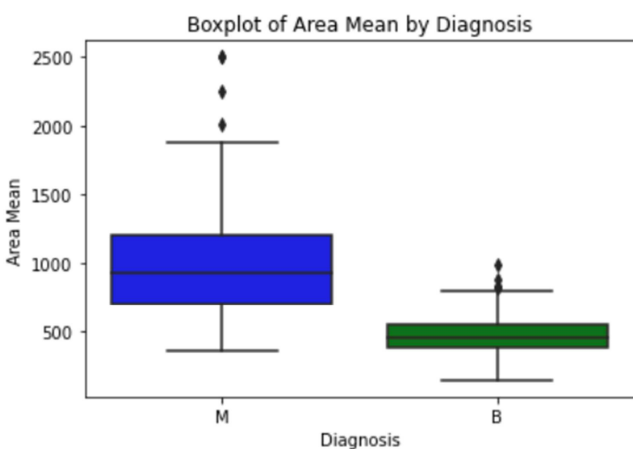


Figure 10. The boxplot below represents the mean area distributions for both diagnosis (blue if for malignancy and green for benign). It shows outliers for M and B.

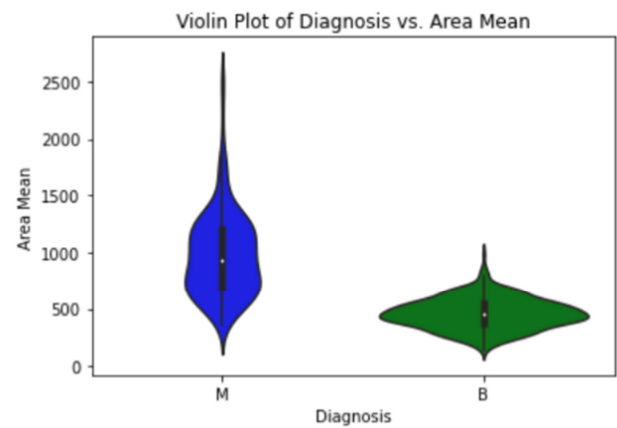


Figure 11. A violin plot h shows peaks in the data. It is used to visualize the distribution of numerical data. Unlike a box plot that can only show summary statistics, violin plots depict summary statistics and the density of each variable. The probability distribution for M shows its peak at 750 and 500 while medians at 1000 and 500 levels.

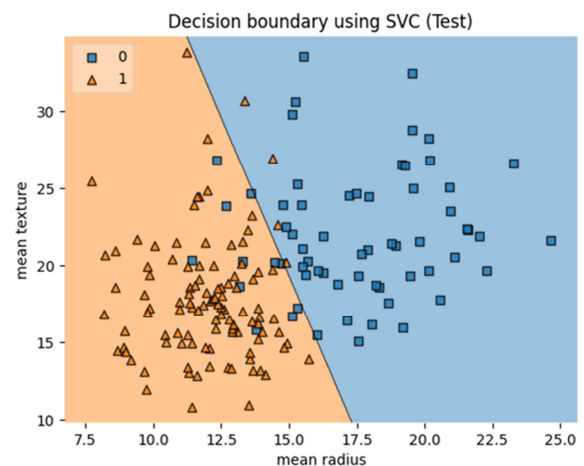


Figure 12. The optimized decision boundary projection on a 2-D plane (mean radius and mean texture).

We have calculated the correlation between the features from the correlation matrix. Such as mean concavity has 92% correlation with mean concave points. Then concavity error has 77% correlation with concave points error. In the statistical analysis the bar graph plot is showing the 0 stands for 'malignant' and 1 stands for 'benign'. (muh-LIG-nunt) A term used to describe cancer.

Malignant cells grow in an uncontrolled way and can invade nearby tissues and spread to other parts of the body

through the blood and lymph system. (beh-NINE) Not cancer. Benign tumors may grow larger but do not spread to other parts of the body. Also called nonmalignant.

Support vector machine and feature engineering.

The classification was done using support vector machine. The decision boundary was found using Lagrange multipliers and Fenchel transform that maps the original problem to a dual problem with convex cost function.

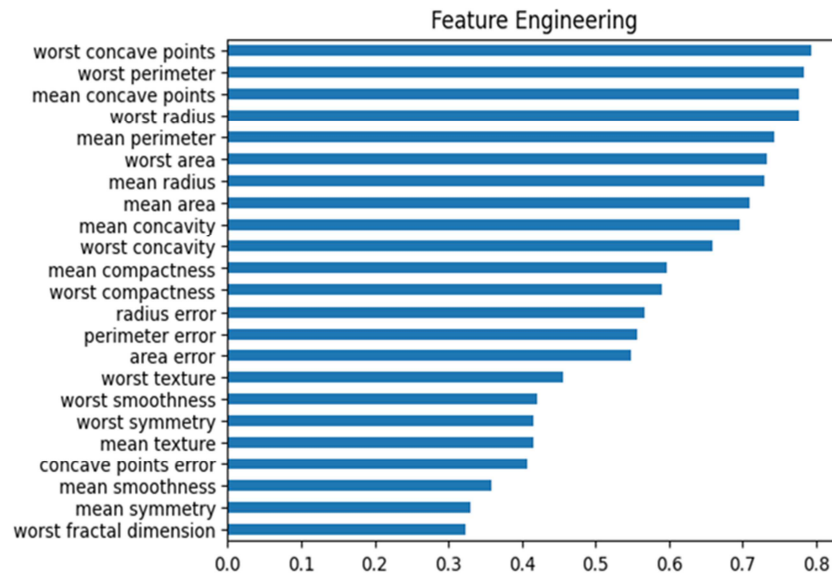


Figure 13. The calculated features that most accurately describe the collected data with respect to binary decision system. M or B.

The numbers at the horizontal line show the importance of each feature in classification.

Let us now perform the classification with different numbers of features.

```
#. Features = 5 : [0.7426900584795322, 0.7622950819672131, 0.8611111111111112]
#. Features = 6 : [0.7485380116959064, 0.7559055118110236, 0.8888888888888888]
#. Features = 7 : [0.7602339181286549, 0.7723577235772358, 0.8796296296296297]
#. Features = 8 : [0.7660818713450293, 0.7833333333333333, 0.8703703703703703]
#. Features = 9 : [0.9181286549707602, 0.9051724137931034, 0.9722222222222222]
#. Features = 10 : [0.9298245614035088, 0.9285714285714286, 0.9629629629629629]
#. Features = 11 : [0.9298245614035088, 0.9285714285714286, 0.9629629629629629]
#. Features = 12 : [0.9532163742690059, 0.9629629629629629, 0.9629629629629629]
#. Features = 13 : [0.9532163742690059, 0.9629629629629629, 0.9629629629629629]
#. Features = 14 : [0.9473684210526315, 0.9714285714285714, 0.9444444444444444]
```

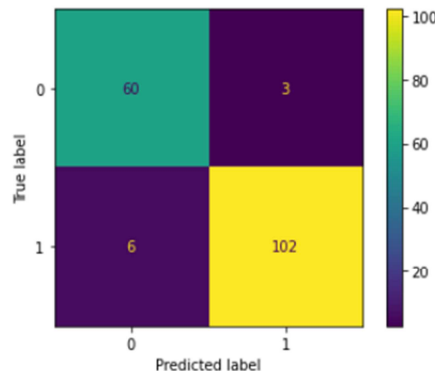


Figure 14. The accuracy of the classification depends on the number of features. The picture shows that the accuracy of the classification Increase to 9.

Data training

We found that the results will depend on the way we split the data on training and test data.

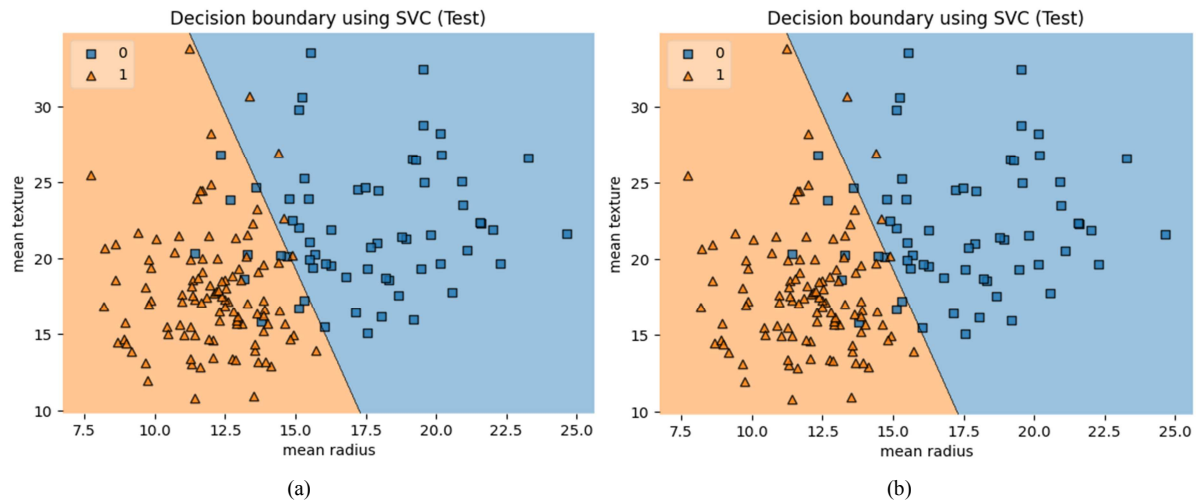


Figure 15. Decision boundary for 70: 30 split (a) and 80: 20 split (b).

Kernel Trick on SVM

We have two classes of observations: malignant tumors and benign tumors the blue points and the purple points. There are numerous ways to separate these two classes as shown in Figure 15. However, we want to find the “best” hyperplane that could maximize the margin between these two classes, which means that the distance between the hyperplane and the nearest data points on each side is the

largest. Depending on which side of the hyperplane a new data point locates, we could assign a class to the new observation.

However, there are a few caveats: not all data are linearly separable. In fact, in the real world, almost all the data are randomly distributed, which makes it hard to separate different classes linearly [14, 16].

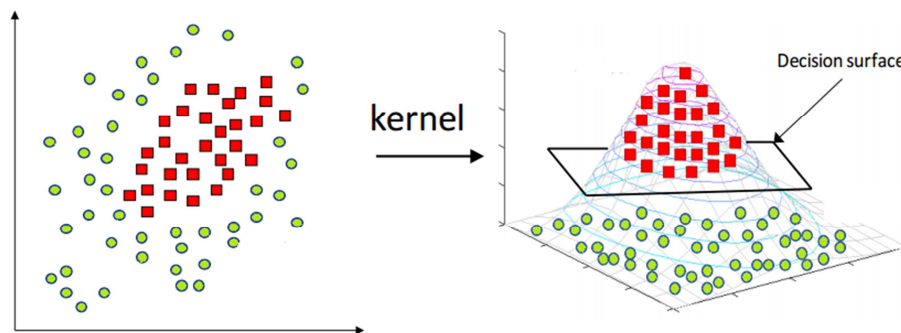


Figure 16. The kernel function maps the point distribution in 2-D to 3-D with more accurate separation.

As one can see in the Figure 16, if we find a way to map the data from 2-dimensional space to 3-dimensional space, we will be able to find a decision surface that clearly divides between different classes. The first thought may be to map all the data points to a higher dimension (in this case, 3 dimensions), find the boundary, and make the classification.

However, when there are more and more dimensions, computations become more and more expensive. This is when the kernel trick comes in.

It allows us to operate in the original feature space without computing the coordinates of the data in higher dimensions.

Let's take a look at the following examples.

```
#RBF Gaussian
svm = SVC(kernel='rbf')

svm.fit(X_train, y_train)

y_pred = svm.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
```

Accuracy: 0.9473684210526315

(a)

```
#Polynomial
svm = SVC(kernel='poly')

svm.fit(X_train, y_train)

y_pred = svm.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
```

Accuracy: 0.9473684210526315

(b)

```
#Linear
svm = SVC(kernel='linear')

svm.fit(X_train, y_train)

y_pred = svm.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
```

Accuracy: 0.9912280701754386

(c)

Figure 17. a) Gaussian Kernel SVM on 9 main features, b) Polynomial Kernel SVM on 9 main features, c) Linear Kernel SVM on 9 main features.

There was an idea of applying kernel trick on SVM, on the best 9 features, to get better accuracy. The RBF Gaussian kernel resulted into 0.94 accuracy;

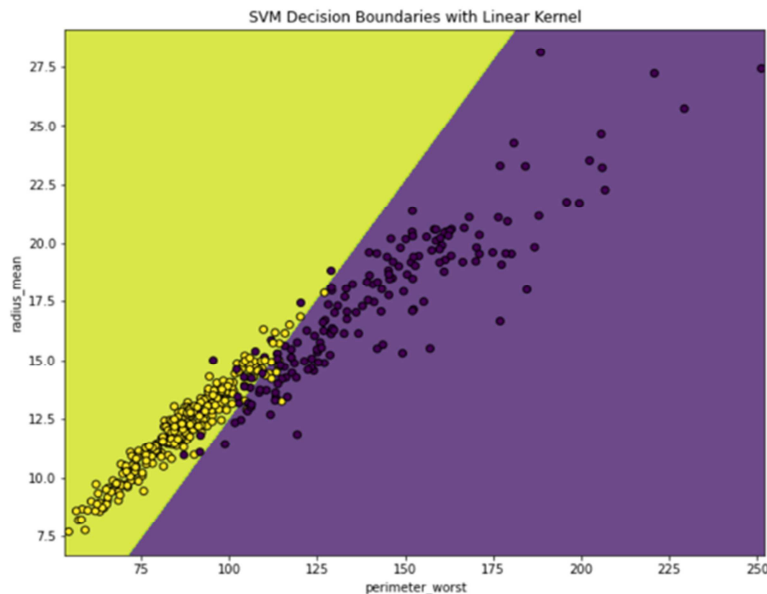
Polynomial resulted into 0.95 accuracy while linear kernel on the 9 features resulted into 0.99 accuracy. Since the 9 features are impossible to simultaneously visualize, we have

visualized features two by two based on the SVM with linear kernel.

Since the correlation does not necessarily determine the accuracy of the model for SVM, we have tried pair by pair, and here are some results visualized below.

Accuracy: 0.9736842105263158

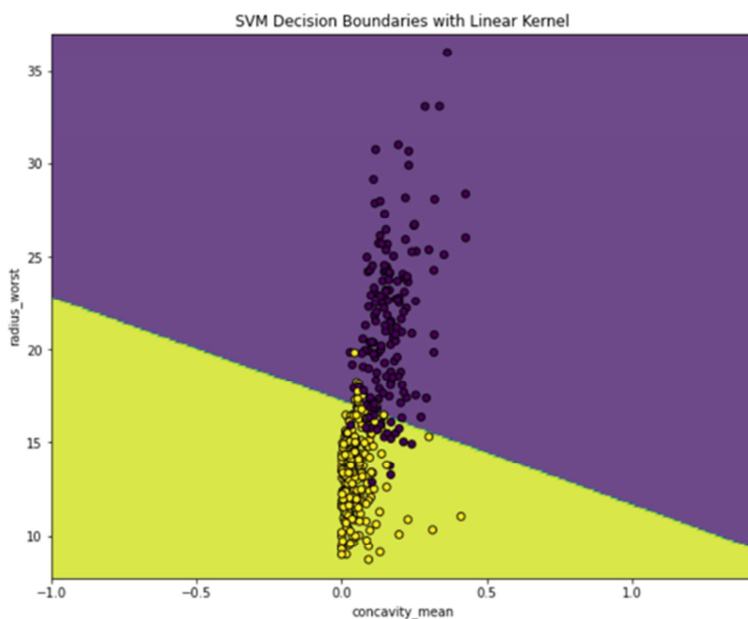
```
/opt/anaconda3/lib/python3.8/site-packages/sklearn/base.py:439: UserWarning: X does not have valid feature names, but
SVC was fitted with feature names
warnings.warn(
```



a)

Accuracy: 0.9473684210526315

```
/opt/anaconda3/lib/python3.8/site-packages/sklearn/base.py:439: UserWarning: X does not have valid feature names, but
SVC was fitted with feature names
warnings.warn(
```



b)

Figure 18. a) Linear Kernel SVM on worst perimeter and mean radius, b) Linear Kernel SVM on worst perimeter and worst concave points.

7. Conclusion

Different machine-learning techniques can be used for the prediction of breast cancer. The challenge is to build accurate and computationally efficient breast data classifiers. In this study, we aimed at analyzing dataset of breast cancer using Support Vector Machine to classify the binary outcome (malignant or benign tumor). [15, 17] We used Kaggle breast cancer data set. The novelty of our approach is we analyzed two algorithms: optimization of the cost function using the hard and soft bound classification expressed in the dual (Lagrange multiplier) form. Moreover, we compared these algorithms with the kernel extension. We compared the performance, the efficiency, and the effectiveness of the models in terms of accuracy, precision, recall, and specificity, to find the best classification accuracy. The results show, that the SVM reaches an accuracy of 99.12% and thus outperforms the other classifiers. Our studies show that out of 30 features only 9 mostly impact the results of the classification. Our current research is focused on the investigation of the performance of a deep learning architecture in performing a breast cancer classification.

References

- [1] Vapnik V. N. Complete Statistical Theory of Learning *Automation and Remote Control*. 80: 1949-1975. DOI: 10.1134/S000511791911002X.
- [2] Vapnik V. N, Lerner A. Y, Chervonenkis A. Y. Learning Methods in Problems of Diagnosis *Ifac Proceedings Volumes*. 2: 741-747. DOI: 10.1016/S1474-6670(17)68922-5.
- [3] Vapnik V. N. An overview of statistical learning theory. *IEEE Transactions On Neural Networks a Publication of the IEEE Neural Networks Council*. 10: 988-99. PMID 18252602 DOI: 10.1109/72.788640.
- [4] Alireza Osareh; Bitia Shadgar, Machine learning techniques to diagnose breast cancer, 2010 5th International Symposium on Health Informatics and Bioinformatics.
- [5] Fan J, Wu Y, Yuan M, Page D, Liu J, Ong IM, Peissig P, Burnside E. Structure-leveraged methods in breast cancer risk prediction. *The Journal of Machine Learning Research*. 2016; 17 (1): 2956–70. [PMC free article]
- [6] Oyewola D, Hakimi D, Adeboye K, Shehu MD. Using five machine learning for breast cancer biopsy predictions based on mammographic diagnosis. *International Journal of Engineering Technologies*. 2016; 2 (4): 142–5. doi: 10.19072/ijet.280563. [CrossRef]
- [7] Hou C, Zhong X, He P, Xu B, Diao S, Yi F, Zheng H, Li J. Predicting Breast Cancer in Chinese Women Using Machine Learning Techniques: Algorithm Development. *JMIR Med Inform*. 2020; 8 (6): e17364. doi: 10.2196/17364. [PMC Free Article]
- [8] Behravan H, Hartikainen JM, Tengström M, Kosma VM, Mannermaa A. Predicting breast cancer risk using interacting genetic and demographic factors and machine learning. *Sci Rep*. 2020; 10 (1): 11044. doi: 10.1038/s41598-020-66907-9. [PMC Free Article]
- [9] Kumar K, Singh VV, Ramaswamy R. Different Perspective of Machine Learning Technique to Better Predict Breast Cancer Survival. *BioRxiv*. 2020 doi: 10.1101/2020.07.03.186890. [CrossRef] [Google Scholar]
- [10] Alghunaim S, Al-Baity HH. On the scalability of machine-learning algorithms for breast cancer prediction in big data context. *IEEE Access*. 2019; 7: 91535–46. doi: 10.1109/ACCESS.2019.2927080. [CrossRef] [Google Scholar]
- [11] Asri H, Mousannif H, Al Moatassime H, Noel T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*. 2016; 83: 1064–9. doi: 10.1016/j.procs.2016.04.224. [CrossRef] [Google Scholar]
- [12] Lotfnezhad Afshar H, Jabbari N, Khalkhali HR, Esnaashari O. Prediction of Breast Cancer Survival by Machine Learning Methods: An Application of Multiple Imputation. *Iran J Public Health*. 2021; 50 (3): 598–605. doi: 10.18502/ijph.v50i3.5606. [PMC Free Article]
- [13] Tapak L, Shirmohammadi-Khorram N, Amini P, Alafchi B, Hamidi O, Poorolajal J. Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. *Clinical Epidemiology and Global Health*. 2019; 7 (3): 293–9. doi: 10.1016/j.cegh.2018.10.003. [CrossRef] [Google Scholar]
- [14] Gaye, B., Zhang, D., & Wulamu, A. (2021). Improvement of support vector machine algorithm in big data background. *Mathematical Problems in Engineering*, 2021, 1–9. <https://doi.org/10.1155/2021/5594899>
- [15] Yamasari, Y., Qoiriah, A., Rochmawati, N., Suartana, I., Putra, O. V., & Nurhidayat, A. I. (2022). *Exploring the kernel on SVM to enhance the classification performance of students' academic performance*. <https://doi.org/10.1109/icvee57061.2022.9930405>
- [16] Shao, J., Liu, X., & He, W. (2021). Kernel Based Data-Adaptive Support vector machines for Multi-Class classification. *Mathematics*, 9 (9), 936. <https://doi.org/10.3390/math9090936>
- [17] *Support Vector machine based diagnosis of breast cancer*. (2020, July 1). IEEE Conference Publication | IEEE Xplore. <https://ieeexplore.ieee.org/abstract/document/9258806>

Biography

Philip de Melo is a data scientist and academic. His research focuses on the development and implementation of new IT technologies including artificial intelligence, machine learning, big data analytics, fast data interoperability, etc. in public health and health care. He was on the faculty of Columbia University (NYC) and Georgia Tech (Atlanta, GA). He served as a PI and Co-PI for a number of projects sponsored by ONR, NSF, AFOSR, ONC, and the industrial project MIDAS.

Mane Davtyan is a bachelor's student at the American University of Armenia focusing on Machine Learning. She is taking an internship class at Armenian Code Academy in the Advanced Machine Learning department and K-Telecom CJSC in the Business Data Analysis department. Despite being new to the field, Mane has already had an opportunity to work with programming languages, like Python and R, and build databases in SQL. Her GitHub profile shows projects about various Machine Learning and Deep Learning models.