DATA SCIENCE

PNEUMONIA DETECTION USING NON-LINEAR SVM

DATA SCIENCE

SUPERVISED CLASSIFICATION OF X-RAY CHEST IMAGES

01 SUPPORT VECTOR MACHINE

Linear and nonlinear

02 DATASET

X ray chest images

03 DATA PROCESSING

Image resize
Image Enhacement
Image Denoising

04 PERFORMANCE MEASURES

Accuracy - Precision
Recall -F1- Score

01

SUPPORT VECTOR MACHINE

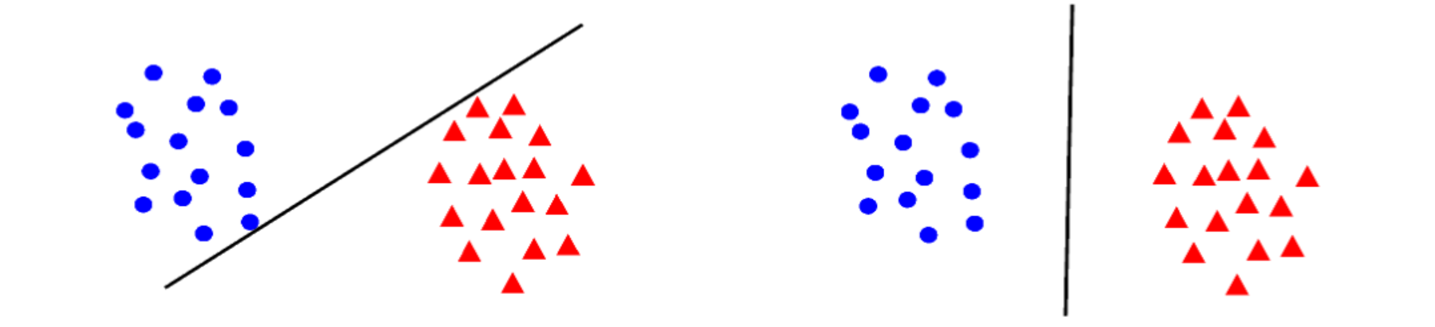
Linear and nonlinear

SVM

Let's assume that A and B are linearly separable sets, meaning there exists a hyperplane H = { x ∈ R^n : w^T x + b = 0 } such that

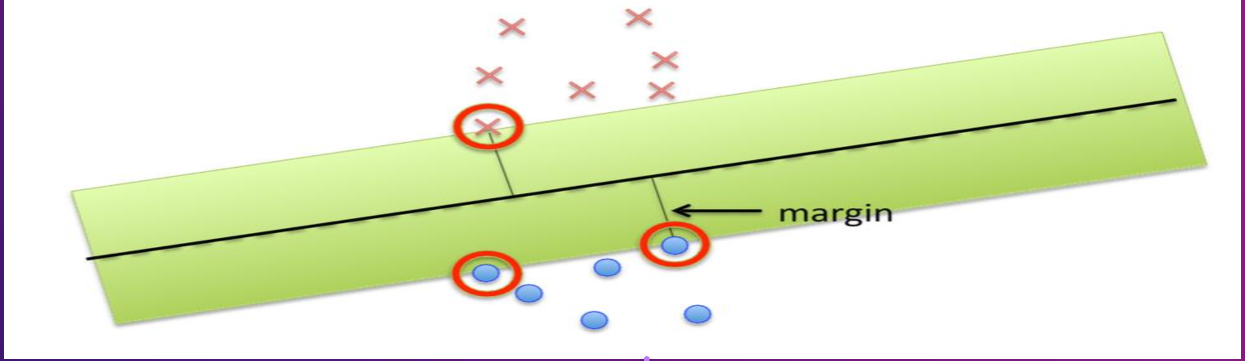
$$\begin{cases} w^T x + b > 0. & y_i = +1 \\ w^T x + b < 0. & y_i = -1 \end{cases} \longrightarrow y_i(w^T x + b) > 0$$

The separating hyperplane is not unique.



SVM GOAL

Find the separating hyperplane that maximizes the distance between the closest training set and the separating hyperplane.



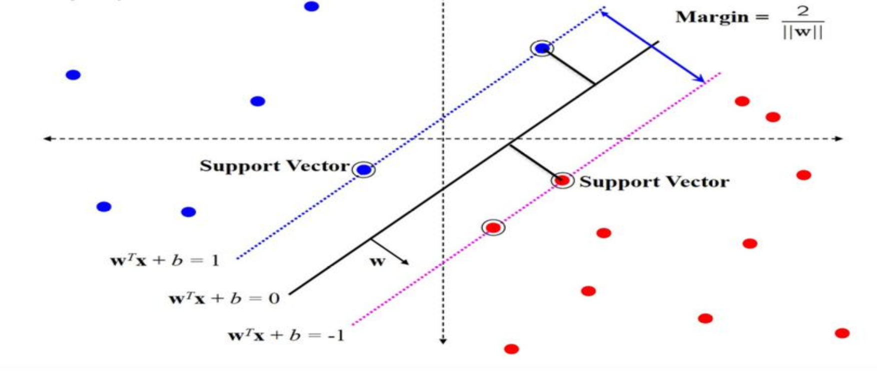
The distance from the closest training point is called the margin.

The circled points are called support vectors.

GEOMETRIC MARGIN

DATA SCIENCE

We define the optimal hyperplane (or maximum margin hyperplane) as:
(w∗, b∗) = argmax ρ(w, b)

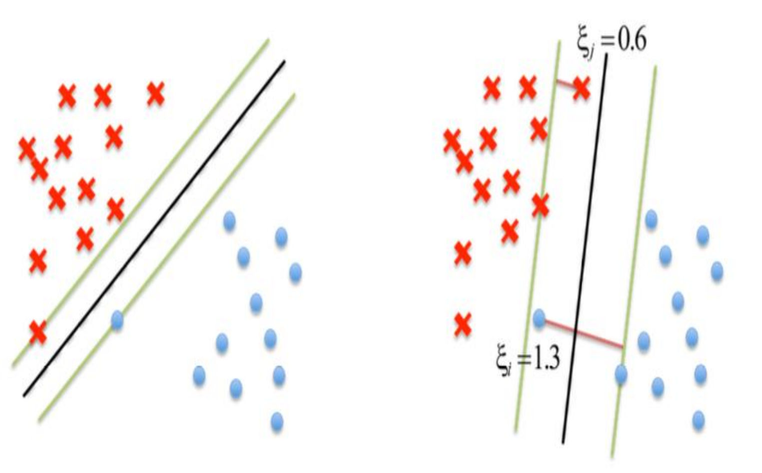

Margin = $\frac{2}{||\mathbf{w}||}$
Support Vector
Support Vector
$\mathbf{w}^T\mathbf{x} + b = 1$
$\mathbf{w}$
$\mathbf{w}^T\mathbf{x} + b = 0$
$\mathbf{w}^T\mathbf{x} + b = -1$

The geometric margin between the two classes is $\frac{2}{\|w\|}$

The larger the margin (separation), the higher the expected generalization

INSIEMI LINEARMENTE NON SEPARABILI

Linear separability is generally too strong an assumption. However, the concept of an optimal separating hyperplane still makes sense, based on slack variables to handle outliers.



The ξi (Slack Variables) account for the non-separability of the data.

$\xi i = 0 \Rightarrow x\_i$ correctly classified (outside the margin)

$\xi i \in (0, 1) \Rightarrow x\_i$ correctly classified (but within the margin)

$\xi i > 1 \Rightarrow x\_i$ misclassified (on the wrong side of the separating hyperplane))

C-SVM

Maximization of the (Soft) Margin and Minimization of the Number of Misclassified Samples

$$\min \frac{1}{2} \parallel w \parallel_2^2 + c \sum_{i=1}^{n} \xi_i$$

$$\text{s.t. } y_i(w^T x + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

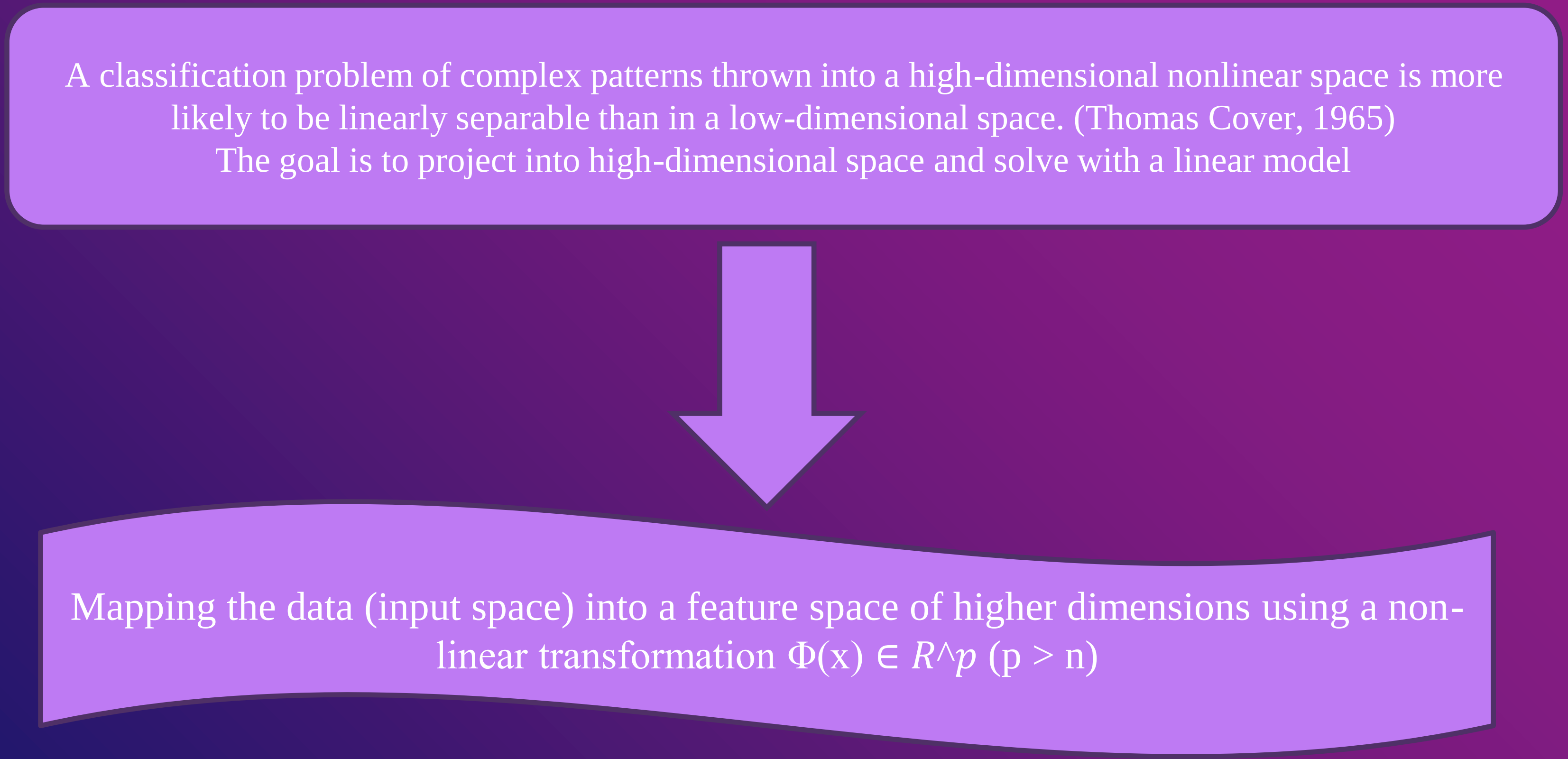
Where C > 0

The regularization parameter C takes into account the penalty for misclassified data.

Larger C⇒ Fewer exceptions (smaller margin, potential overfitting).

Smaller C⇒ more eccezioni exceptions (larger margin, potential underfitting).

Non linear SVM

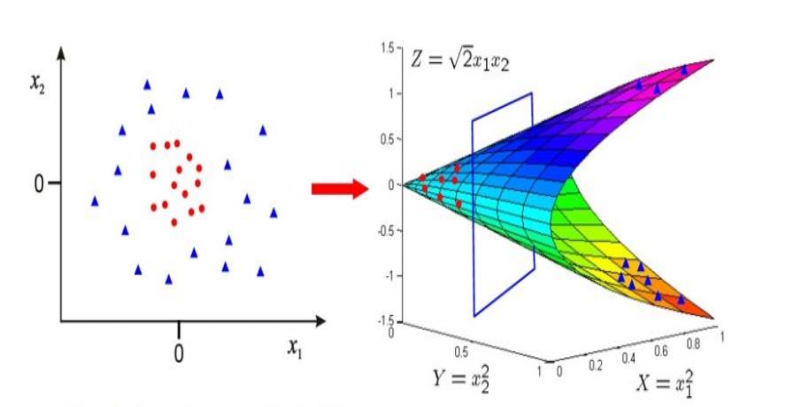


A classification problem of complex patterns thrown into a high-dimensional nonlinear space is more likely to be linearly separable than in a low-dimensional space. (Thomas Cover, 1965)
The goal is to project into high-dimensional space and solve with a linear model

Mapping the data (input space) into a feature space of higher dimensions using a non-linear transformation $\Phi(x) \in R^p$ $(p > n)$

KERNEL

Applying SVM to Φ(x_i) rather than to x_i, such that the formulation of SVM (in the feature space) is expressed only using the inner product ⟨Φ(x_i), Φ(x_j)⟩. To achieve this, it is sufficient to know how to compute the scalar product k(x_i, x_j) in the feature space.
k is called a kernel, and the same kernel can correspond to different feature maps Φ.



The idea of the kernel function is to perform operations in the input space rather than in the potentially high-dimensional feature space. Therefore, the inner product does not need to be evaluated in the feature space. We want the function to map observations from the input space to the feature space.

02 DATASET

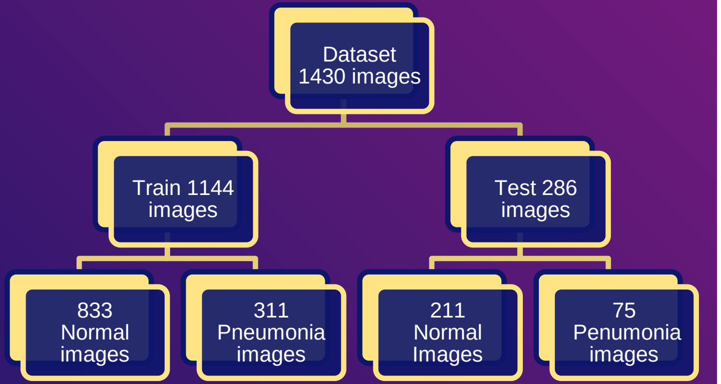
X ray chest images

DATASET

The dataset is organized into 2 different types (train and test) and contains various categories of images (Pneumonia/Normal). There are 1,430 chest X-ray images (JPEG) and 2 categories (Pneumonia/Normal).

For the analysis of chest X-ray images, all chest X-rays initially underwent a quality check, eliminating all low-quality or unreadable scans. The diagnoses of the images were then classified by two expert physicians before being authorized for model training.

DATASET SPLITTING


Dataset 1430 images

Train 1144 images

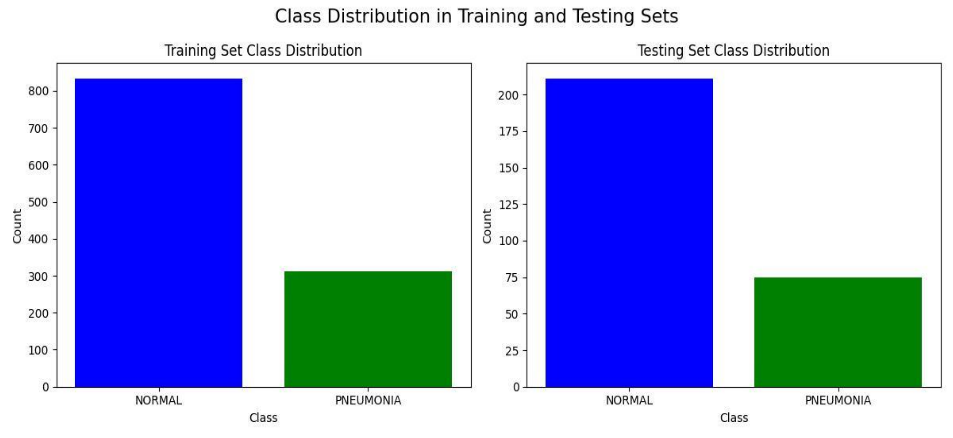
Test 286 images

833 Normal images

311 Pneumonia images

211 Normal Images

75 Penumonia images


Class Distribution in Training and Testing Sets

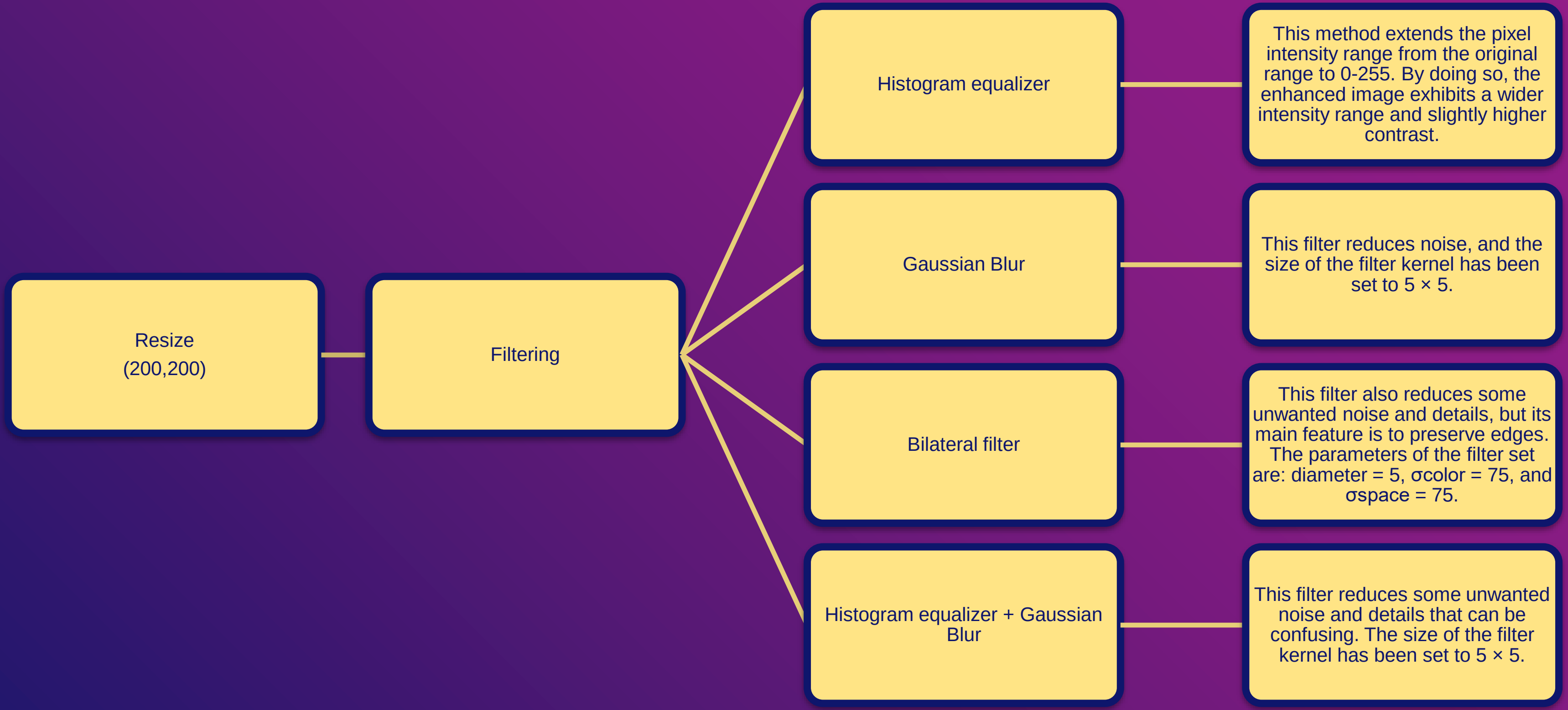
03

DATA PROCESSING

Image resize
Image Enhacement
Image Denoising

PREPROCESSING DATASET

DATA SCIENCE



Resize
(200,200)

Filtering

Histogram equalizer

This method extends the pixel intensity range from the original range to 0-255. By doing so, the enhanced image exhibits a wider intensity range and slightly higher contrast.

Gaussian Blur

This filter reduces noise, and the size of the filter kernel has been set to 5 × 5.

Bilateral filter

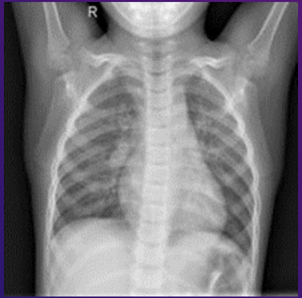
This filter also reduces some unwanted noise and details, but its main feature is to preserve edges. The parameters of the filter set are: diameter = 5, σcolor = 75, and σspace = 75.

Histogram equalizer + Gaussian Blur

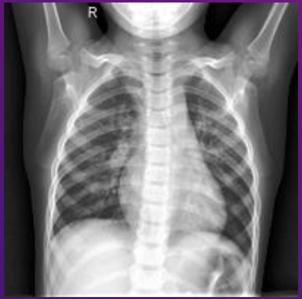
This filter reduces some unwanted noise and details that can be confusing. The size of the filter kernel has been set to 5 × 5.

RESIZING AND FILTERING

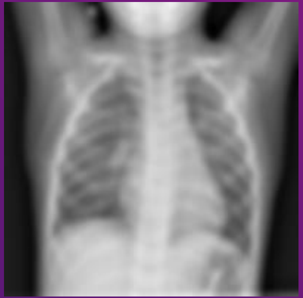
Immagine Resized



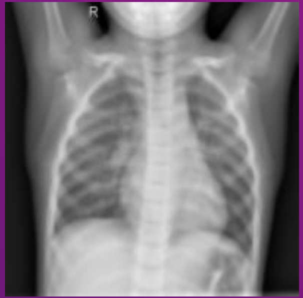
Histogram equalizer



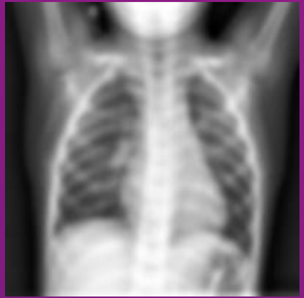
Gaussian Blur



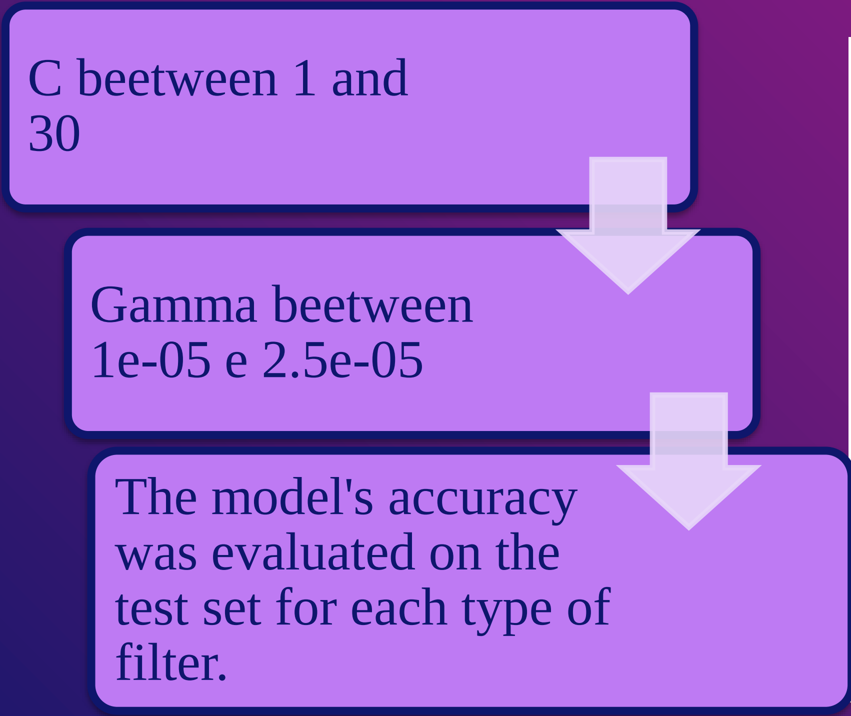
Bilateral Filter



Histogram equalizer + Gaussian Blur



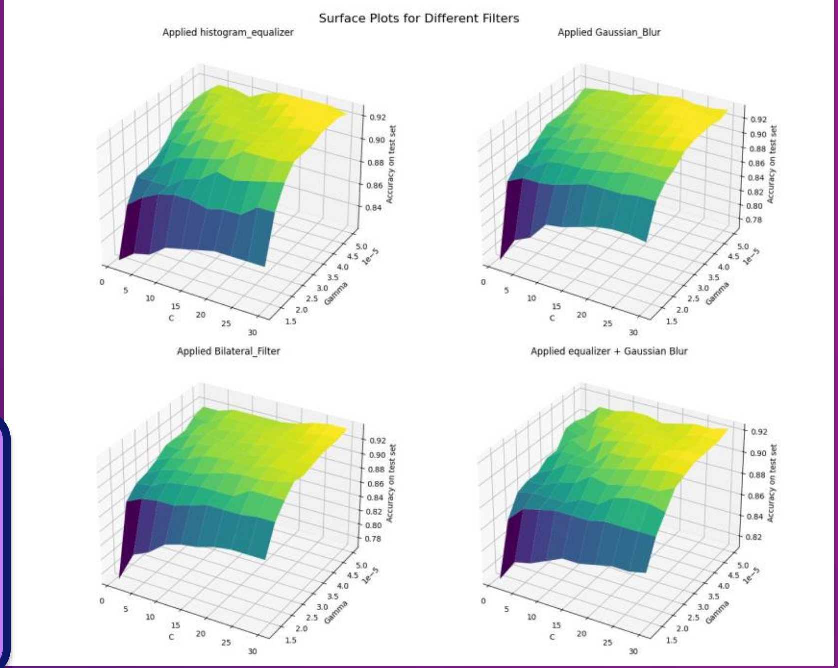
STIMA DEI PARAMETRI C E GAMMA



C beetween 1 and 30

Gamma beetween 1e-05 e 2.5e-05

The model's accuracy was evaluated on the test set for each type of filter.



04

PERFORMANCE MEASURES

Accuracy – Precision - Recall - F1- Score

PERFORMANCE MEASURES

Accuracy

- Accuracy represents the percentage of correct predictions compared to the total predictions.

Precision

- Precision indicates the percentage of correct positive class predictions compared to the total positive class predictions.

Recall

- The percentage of correct positive class predictions compared to the total positive cases.

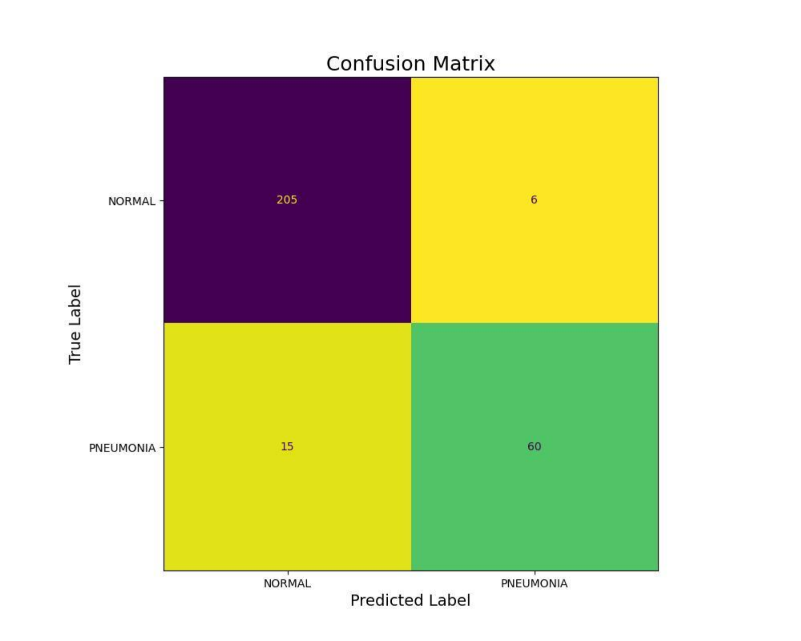
F1-Score

- The harmonic mean of Precision and Recall

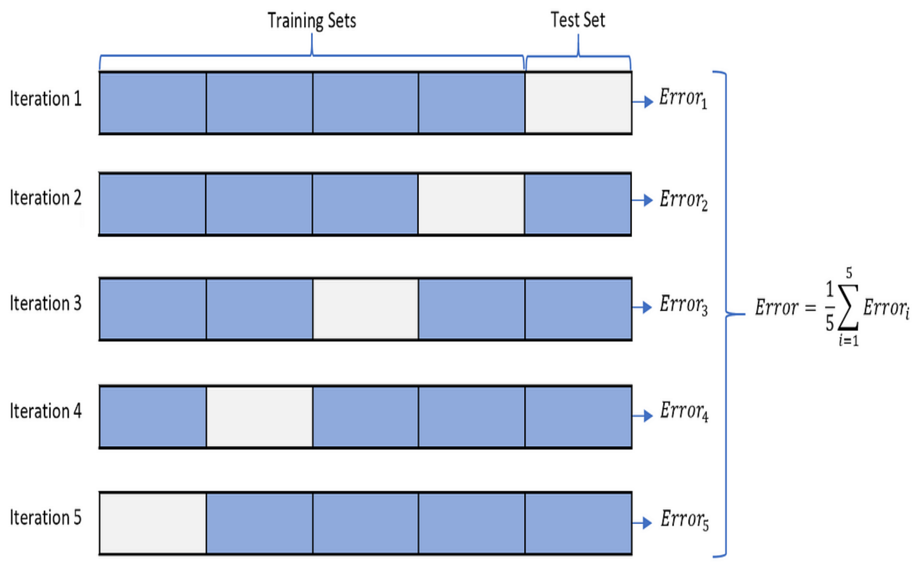
RESULTS

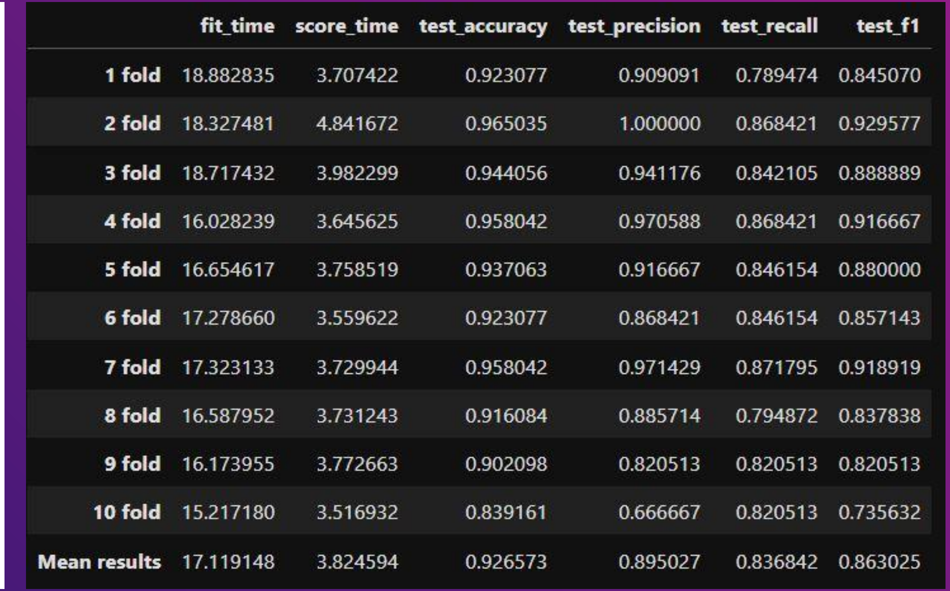
Results of the model with:

- **Bilateral filter**
- **C = 30**
- **Gamma = 4.58e-05**
- Accuracy = 0.93
- Precision = 0.91
- Recall = 0.8
- F1 score = 0.85



10 FOLD CROSS VALIDATION RESULTS





	fit_time	score_time	test_accuracy	test_precision	test_recall	test_f1
1 fold	18.882835	3.707422	0.923077	0.909091	0.789474	0.845070
2 fold	18.327481	4.841672	0.965035	1.000000	0.868421	0.929577
3 fold	18.717432	3.982299	0.944056	0.941176	0.842105	0.888889
4 fold	16.028239	3.645625	0.958042	0.970588	0.868421	0.916667
5 fold	16.654617	3.758519	0.937063	0.916667	0.846154	0.880000
6 fold	17.278660	3.559622	0.923077	0.868421	0.846154	0.857143
7 fold	17.323133	3.729944	0.958042	0.971429	0.871795	0.918919
8 fold	16.587952	3.731243	0.916084	0.885714	0.794872	0.837838
9 fold	16.173955	3.772663	0.902098	0.820513	0.820513	0.820513
10 fold	15.217180	3.516932	0.839161	0.666667	0.820513	0.735632
Mean results	17.119148	3.824594	0.926573	0.895027	0.836842	0.863025

Thanks for your attention!

OUR TEAM

GIANMARCO BORRATA

PASQUALE PIPICIELLO