Bare Demo of IEEEtran.cls for IEEE Computer Society Journals

Michael Shell, *Member, IEEE,* John Doe, *Fellow, OSA,* and Jane Doe, *Life Fellow, IEEE*

Abstract—The abstract goes here.

Index Terms—Computer Society, IEEE, IEEEtran, journal, LATEX, paper, template.

1 INTRODUCTION

2 METHODOLOGY

As the amount of generated per day grows at an exponential rate, brand new technologies have to be developed to cope up with the copius exabytes of data. Machine learning tools provide us with the capabilities to handle both structured and unstructured datasets. These tools can be configured to analyze patterns inherent in the data and make accurate predictions based on the information obtained. This concept is a reality for almost all sectors today. As per a 2020 Stanford study, the amount of healthcare data generated will be around 2,314 exabytes with a steady growth of 48%. The pipeline developed for this project has been depicted in Figure 1. The remainder of the section descibes each individual step in detail.

2.1 Obtaining Raw data

In order to obtain distinct yet comparable subjects, a cohort dataset of 89 patients was selected in this study. The dataset consisted of four intrinsic molecular subtypes of breast cancer which are contrasted on the genes a cancerous cell expresses. The dataset has been descibed in Table 1.

For each of the patient, a CT scan was conducted to obtain cross-sectional images of the hypothesised tumor location. CT scans provide a more detailed description of the patients condition by increasing the radiation level the patient is exposed to. Once the scan is completed three views are obtained namely, Axial, Sagittal and Coronal. A sample Axial view has been displayed in Figure 2 with the distinct grey circular mass on the right depicting the tumor. DICOM (Digital Imaging and Communication in Medicine) images were obtained after the scan. For each patient 323 new studies were conducted with each study have 384 series which corresponded to 466 instances or images of the scan. Even though DICOM files are a standard format for medical imaging, NRRD (Nearly Raw Raster Data) files are anonymmized and contain no sensitive patient information. Moreover NRRD store the entire information in a single file as opposed to DICOM imaging.

2.2 Convert to a suitable format

As mentioned previously, NRRD provides a more insightful appraoch to understanding medical imaging and recognizing inherent patterns in a concised format. The conversion was done with the help of the Plastimatch tool which is an open source software for image computation. Plastimatch takes the DICOM image which is described in a polyline vectorized format, and converts it into a series of pixels which is more prominently known as rasterization. The

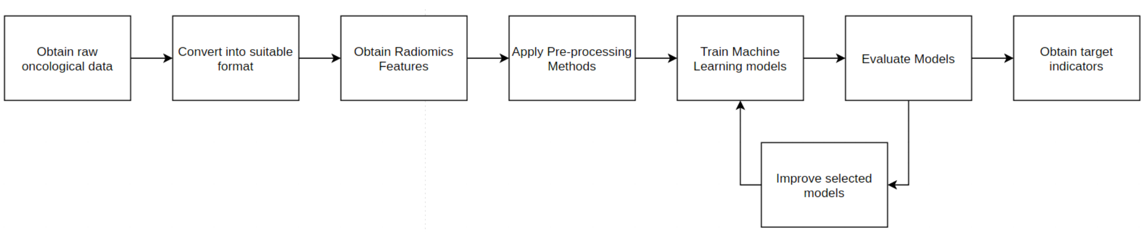

Fig. 1. Project Pipeline

TABLE 1
Data Description

Subtype	Number of paitents	Estrogen Receptor	Progesterone Receptor	HER2	KI67 range
Luminal A	29	+	+/-	-	[5,20]
Luminal B	36	+	+/-	+/-	[25,80]
Triple Negative	19	-	-	-	[20,90]
HER	5	-	-	+	[30,50]

subroutine for rasterization of a DICOM image set with coordiantes x and y is shown below.

```
def rast(x, y, shape):
        nx, ny = draw.polygon(x, y, shape)
        nrrd = np.zeros(shape, dtype=np.bool)
        nrrd[ny, nx] = True

        return nrrd
```

Once this step is conducted, our image is in a compressed format, rife with information. Information extraction can be conducted through multiple means such as using neural networks, OCR recognition or pattern recognition algorithms.

2.3 Obtaining Radiomics Features

Information extraction from images directly has certain drawbacks. For eg, consider tumor classification using a standard Convolutional Neural Network (CNN). The CNN might be extremely successful in determining the existense of a blob of mass and it's exact location. However diagnosing the exact nature and feature set of the tumor is extremely difficult for a CNN. This is because a CNN views the image as simply a collection of pixels without any regard to the information embedded in all the views of the data.

To tackle this issue, we have utilized radiomics algorithms to extract feature sets from the medical images to reveal characteristics which are not captured by trained networks. The open-source Python library, PyRadiomics was used to mine out the required feature set. Before the actual extraction could be performed, a set of filters were applied on the NRRD to provide a comprehensive view of the data. The filters applied are listed in table 2.

To define a Region of Interest (ROI) and to check the dimensional constrainsts of the data, a mask file is utilized. The mask image corresponding to Figure 2 is shown in Figure 3. Note the red mark demarcating the tumor is done by a radiologist as is standard procedure. The features are now extracted from the image set with the help of the mask file. The features extracted are descibed by the Imaging Biomarker Standardization Initiative (IBSI). The features have been shown in tables 3 and 4.

Therefore for each patient, the total number of features obtained are number of filters $\times$ number of features i.e, 17 $\times$ 100 = 1700 features. Now that the entire feature set has been collected, we can begin the classification task.

2.4 Applying Pre-processing Techniques

From the 1700 features collected, not all of the features will contribute equally in the classification function. The process of preparing the input data for pattern learning by removing redundant characteristics, reducing noises and normalizing, selecting, and extracting features is termed as Data Pre-Processing.

3 Conclusion

The conclusion goes here.

Appendix A
Proof of the First Zonklar Equation

Appendix one text goes here.

Appendix B

Appendix two text goes here.

Acknowledgments

The authors would like to thank...

References

[1] H. Kopka and P. W. Daly, *A Guide to LATEX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

Fig. 2. Axial view with the tumor

TABLE 2
Applied Filters

Filter	Description	Equation
Wavelet (9)	Selective emphasizing or de-emphasizing of image in selected spatial frequency domain	-
Square	Square the image intensities	x := $(cx)^2$
Square Root	Compute root of image intensities	x := $\sqrt{cx}$
Laplacian of Gaussian σ = 1, 2, 3	Applies a Laplacian of Gaussian filter to the input image and yields a derived image for each sigma value specified	$\frac{1}{(\sigma\sqrt{2\pi})^3} e^{-\frac{x^2+y^2+z^2}{2\sigma^2}}$
Logarithm	Computes the natural logarithm of image intensities	$c\log(x+1)$
Exponential	Computes the exponential of the original image	e^{cx}
Gradient	Computes the gradient of the image	-

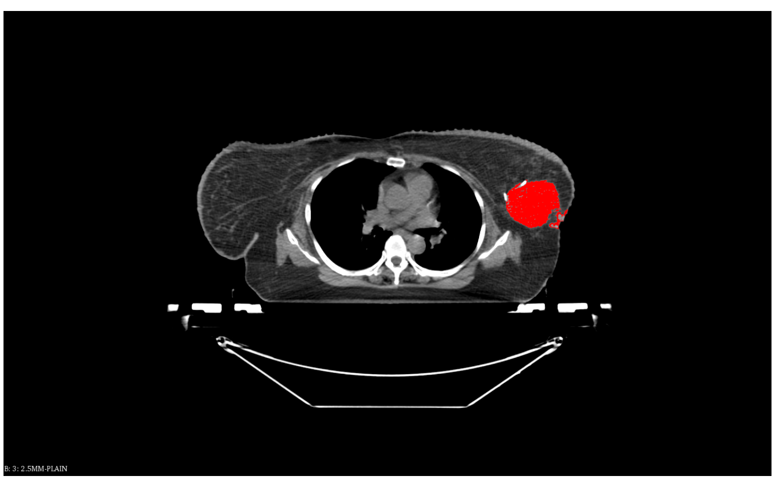

Fig. 3. Mask image

TABLE 3
Features-I

Feature Class	Feature	Feature Class	Feature
Shape	Elongation	Grey Level Co-occurance Matrix	Autocorrelation
	Flatness		Cluster_Prominence
	Least_Axis_length		Cluster_Shade
	Major_Axis_Length		Cluster_Tendency
	Max_2D_Diameter_Column		Constrast
	Max_2D_Diameter_Row		Correlation
	Max_2D_Diameter_Slice		Difference_Average
	Max_3D_Diameter		Difference_Entropy
	Mesh_Volume		Difference_Variance
	Minor_Axis_Length		Inverse_Variance
	Sphercity		Joint_Average
	Surface_Area		Joint_Energy
	Surface_Volume		Joint_Entropy
	Voxel_Volume		MCC
			Maximum_Probability
			Sum_Average
			Sum_Entropy
			Sum_Squares
			Id
			Idm
			Idn
			Idmn
			Imc1
			Imc2
First Order Statistics	10 Percentile	Grey Level Run Length Matrix	Uniformity
	90 Percentile		Normalized_Uniformity
	Energy		Variance
	Entropy		High_Run_Emphasis
	Interquartile_Range		Long_Run_Emphasis
	Kurtosis		Long_High_Run_Emphasis
	Maximum		Long_Low_Run_Emphasis
	Mean_Absolute_Deviation		Low_Run_Emphasis
	Mean		Run_Entropy
	Median		Run_Uniformity
	Minimum		Run_Uniformity_Normalized
	Range		Run_Percentage
	Robust_Mean_Deviation		Run_Variance
	Robust_Mean_Squared		Short_Run_Emphasis
	Skewness		Short_Run_High_Emphasis
	Total_Energy		Short_Run_Low_Emphasis
	Uniformity		
	Variance		

TABLE 4
Features-II

Feature Class	Feature
Grey Level Size Zone Matrix	Non_Uniformity Non_Uniformity_Normalized Variance High_Zone_Emphasis Large_Area_Emphasis Large_Area_High_Level_Emphasis Large_Area_Low_Level_Emphasis Low_Zone_Emphasis Zone_Non_Uniformity Zone_Non_Uniformity_Normalized Small_Area_Emphasis Small_Area_High_Level_Emphasis Small_Area_Low_Level_Emphasis Zone_Entropy Zone_Percentage Zone_Variance
Gray Level Size Zone Matrix	Dependence_Entropy Dependence_Non_Uniformity Dependence_Non_Uniformity_Normalized Dependence_Variance GL_Non_Uniformity GL_Variance High_Emphasis Large_Dependence_Emphasis Large_Dependence_High_Emphasis Large_Dependence_Low_Emphasis Low_Emphasis Small_Dependence_Emphasis Small_Dependence_High_Emphasis Small_Dependence_Low_Emphasis
Neighbouring Gray Tone Difference Matrix	Busyness Coarseness Complexity Constrast Strength