

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, L^AT_EX, 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 copious 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 describes 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 described 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 anonymized 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 approach 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

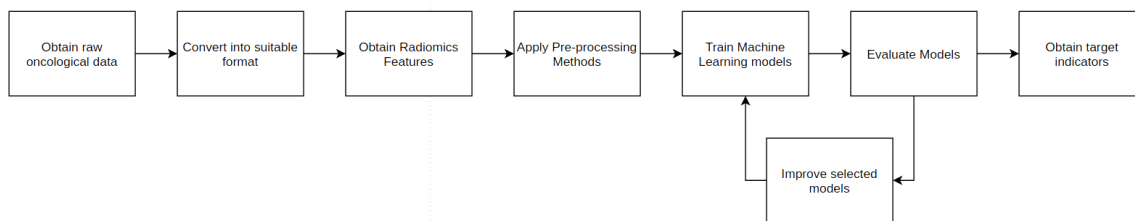


Fig. 1. Project Pipeline

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 constraints of the data, a mask file is utilized.

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	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	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	$\text{clog}(x + 1)$
Exponential	Computes the exponential of the original image	e^{cx}
Gradient	Computes the gradient of the image	-