DigiOnco: A Pipeline to Unveil Digital Non-Invasive Biomarkers from Multi-parametric Radiomics Footprints

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

1.1 Personalized Therapeutics: An oveview

Personalized medicine, diagnostics and therapeutics upholds the promise of accurate decision making by leveraging the power of machine learning/deep learning based Digitized Imaging and Anlytical Models (DIAM). An unprecendented paradigm shift is trending in the ease with which the medical fraternity embraces DIAMs. Promising collaborations have strengthened interactions among medical specialists and technological research groups, resulting in automated and reproducible analytics, more specific in the area of oncological research. The higher levels of concordance, ability to deduce image patterns not visible to the trained human eye, along with reduced intra and interobserver variations and subjectivity, have emerged as promising catalysts for embracing DIAMs.

1.2 Tumor Heterogenity in Onological Research: The Real Challenge for DIAM

Tumor heterogenity, involving a wide range of morphological phenotypes and prognostic variables, has been a great challenge in oncological diagnostics and prognosis. For example, in the much studied vertical of breast carcinoma, therapeutic decision making involves multi-modal analytics, including characterizing the morphology and grading a tumor, histopathology, immunohistochenomistry (IHC) and insitu hybridization (ISH). The biomarkers thus obtained, are evaluated clinically and analytically for their optimal clinical application. The profiling is further guided by the higher rate of concordance and robustness. With the multiparametric molecular assays being very expensive, approximate mutigene testing and surrogate definitions of intrinsic subtypes can be arrived at using IHC measurments.

1.3 Radiomics Assisted Digitization: An aid for Oncological Decision Making admist heteroginity

With the wide-spread know how of computed tomography (CT), positron emission tomography (PET), and magnetic resonance imaging (MRI) imaging methods, Radiomics gathered much attention in the last few years of oncological research. Radiomics pipelines study the quantitative features of the image under consideration, by extracting the first-order, second-order and higher order statistical features of the Region of Interest (ROI). The hypothesis states that when a simultaneous study of heterogenous groups of parameters of a single lesion is performed, a filtered, appropriate and customized subset of parameters (called digital biomarkers) across groups might emerge. These digital biomarkers which define specific indicative tissue characteristics, when combined with the clinical biomarkers, have the long-standing potential to offer personalized therapeutics to the patient.

1.4 DIAM for Oncology: State-of-the-Art

The stability and reproducability of the existing Radiomics model, along with a need to standardize the assessment of digital biomarkers and cross-validation techniques, are indeed a matter that needs immediate attention before including them in the diagnostic routine. Moreover, statistical associations are, to a greater extent, confounded by the patient centric parameters like age, sex, habits, phenotypes and genotypes, that can have a profound impact on the model performance. Existing Radiomics models, as presented in Table are predominantly applied to MRI imaging, due to it's monochrome image quality and wide availability of literature in terms of statistical image analytics.

1.5 DigiOnco: A Pipeline to Unveil Digital Non-Invasive Biomarkers from PET/CT scans

In this paper, we present DigiOnco, a novel pipeline, intricately woven with carefully chosen set of algorithms. DigiOnco unveils the digital non-invasive biomarkers from multi-parameteric Radiomics footprints obtained from the PET/CT imaging techniques. The hypothesis generation and validation is performed as both internal cross-validation as well as a retrospectively validated study on an independent cohort, having a set of external and independent group of patients.

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. 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 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.

PyRadiomics obtains radiomics features from the CT scan results in a stagewise manner. Initially the images are loaded into the platform by using SimpleITK which supports a gamut of image types along wit basic image processing techniques. In the next step, the filters descibed in 2 are applied using SimpleITK, PyWavelets, and Numpy.

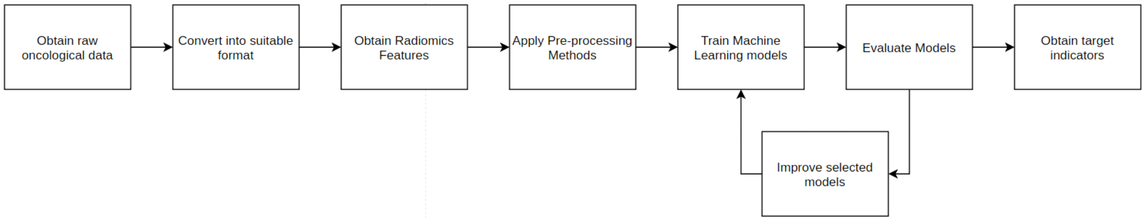
Obtain raw oncological data
Convert into suitable format
Obtain Radiomics Features
Apply Pre-processing Methods
Train Machine Learning models
Evaluate Models
Obtain target indicators
Improve selected models

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(TN)	19	-	-	-	[20,90]
HER	5	-	-	+	[30,50]

Finally, statistical and texture classes are used for feature extraction. The features so obtained, are stored in a dictionary format which suitable labels.

To define a Region of Interest (ROI) and to check the dimensional constrainsts of the data, a mask file is utilized. The mask file contains the tumor's location demarcated by a radiologist. The features extracted are descibed by the Imaging Biomarker Standardization Initiative (IBSI) and have have been shown in tables 3 and 4.

Therefore for each patient, the total number of features obtained are number of filters × number of features i.e, 17 × 100 = 1700 features. Once the entire feature set has been collected, the classification task can be started.

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. Multiple data pre-processing techniques have been applied to the feature set. These techniques have been descibed in Table 5.

Since the number of test subjects for each class is not similar, a threshold confidence level must be specified during the hypothesis testing phase. A 'P-value' is utilized in hypothesis testing to test the hypothesis under observation. A lower p-value corresponds to a higher confidence level in the predictions. The number of features selected after the pre-processing step is directly proportional to the p-value as a higher p-value will be more accomodating of even unimportant features. A grid for different p-values was created and the corresponding number of features were obtained.

2.5 Model-based Predictions

Once the features have been narrowed down, the model building process begins. For any task on hand, we have a wide array of classifiers which accurately predict the nature of the test set. The set of classification algorithms considered are shown in Table 6. In order to determine which algorithm would perform the best for our cohort dataset, we trained all the models on a standard benchmark dataset belonging to the same field i.e, the Winconsin Breast Cancer Diagnostic Dataset. The tabulated results for each algorithm is shown in Table 7.

As determined, SFORCE (post validation) provides promising results without overfitting and hence is used to classify test subjects into the target classes. SFORCE establishes a symbiotic relation between a predictive model (Random Forest) and an Ensemble model (AdaBoost). Both these models work on the presented data simlutaneously, aiding each other in the prediction process. Random Forests provides a strong learning system with the occasional pitfall

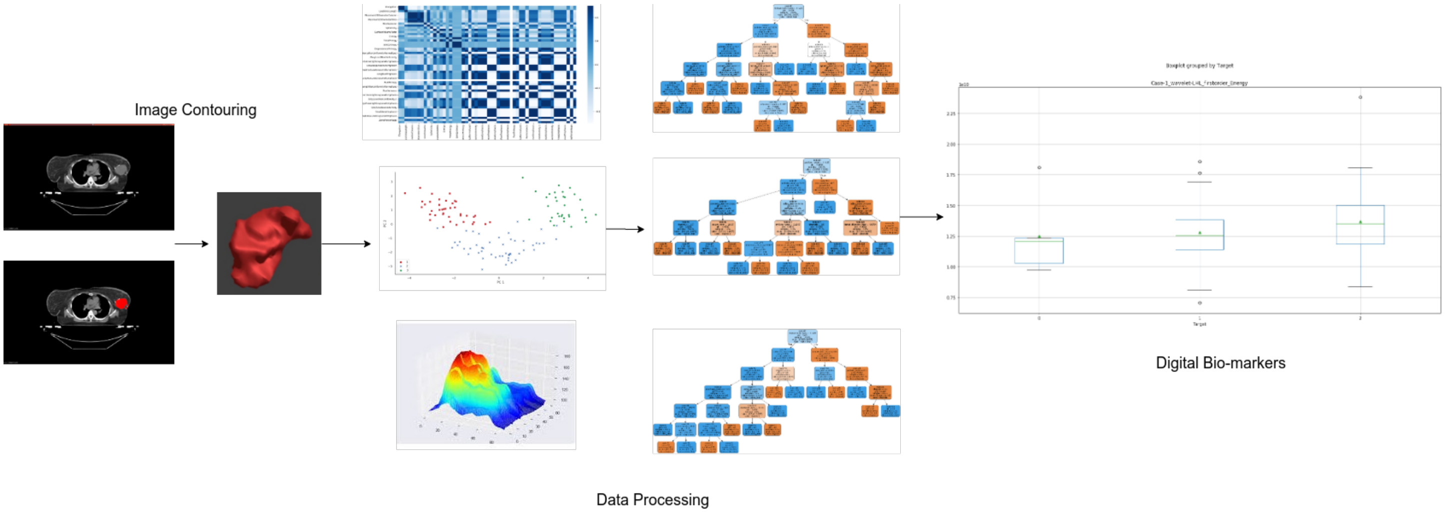
Image Contouring
Boxplot grouped by Target
Case-1_wavelet-LHL_firstorder_Energy
Target
Digital Bio-markers
Data Processing

Fig. 2. untitled

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

of overfitting. The data is classified based the features which contrast the classes with the highest information content. The process of data classification using Random Forest is shown in Algorithm 1. AdaBoost solves the problem of overfitting by presenting the system with the misclassified data and forcing it to improve the overall performance. The two flavours of AdaBoost i.e, SAMME and SAMME.R have been descibed in Algorithms 2 and 3. SFORCE combines the strength of Random Forests and takes care of the drawbacks by using a Boosting algorithm to make the search process more concentrated as shown in Algorithm 4.

To obtain digital bio-markers, two cases studies were conducted from the avaiable cohort dataset. The first study involved classifying test subjects as TN or non TN subjects. In the second study, the Luminal-B dataset was set aside as the test dataset due to the close resemblance of it's characteristics with those of Luminal A. The model was trained to place the test subjects into the Luminal-A class with an accuracy of 72.7%. The results for different p-values have been descibed in Tables 8 and 9. Based on these results, box-plots have been obtained for the selected features which act as bio-markers for future reference.

Algorithm 1 : Ensemble Learning: Random Forest

```
// Input: Data Set D = $\{(x_1, y_1)$,$(x_2, y_2), \dots \; ((x_m, y_m)\}$, Feature Set F, Randomization Factor R, Number of trees T
// Output: Root node of i^th tree
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
for $\forall i \in \{1, 2, \dots T\}$ do
    $N_i$ ← Root node of i^th tree
    if All targets belong to same class i.e $y_i$ or $F \in \emptyset$ then
        Return $N_i$
    end if
    $D_i$ ← bootstraped sample from D
    for Each node do
        $f$ ← Randomly selected $R$ features from $F$
        $N_f$ ← Best Feature from $f$ features
        $N_p$ ← Best Split based on $N_f$
    end for
end for
return $N_i$
```

3 Results and Conclusion

From the data-driven pipeline, quantifiable digital biomarkers were obtained in the form of box and whisker plots. These plots provide a convinient method of displaying the data distribution and provide insight to the oncological expert during prognosis of future test subjects. Sample box

TABLE 3
Features-I

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

Algorithm 2 : Stagewise Additive Modeling: SAMME

```
// **Input:** Data Set D = {$(x_1, y_1)$,$(x_2, y_2)$, ... ($(x_m, y_m)$}, Number of Learning Rounds T, Learning Algorithm $\epsilon$
// **Output:** sign($\sum_{t=1}^{T} \alpha_t.C_t$)
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
$D_1(x)$ = 1/m {Initialize the weight distribution}
**for** $t = \{1, 2, \ldots T\}$ **do**
    $C_t$ = $\epsilon$(D,$D_t$) {Create classifier $C_t$}
    $e_t$ = $P_{x\sim \mathrm{D}}(h_t(x) \neq f(x))$ {Calculate error $e_t$}
    $\alpha_t = log\frac{1 - e_t}{e_t}$ + log($K$-$1$) {Calculate the weight $h_t$}
    $D_i(x) \leftarrow D_i(x)$.exp($\alpha_t$.P($C_i \neq f(x)$)) {Update the distribution $D_t$}, $i = \{1, 2, \ldots m\}$
    Renormalize $D_t(x)$
**end for**
```

Algorithm 3 : Stagewise Additive Modeling for Real Value Predictions: SAMME.R

```
// **Input:** Data Set D = {$(x_1, y_1)$,$(x_2, y_2)$, ... ($(x_m, y_m)$}, Number of Learning Rounds T, Learning Algorithm $\epsilon$
// **Output:** sign($\sum_{t=1}^{T} \alpha_t.C_t$)
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
$D_1(x)$ = 1/m {Initialize the weight distribution}
**for** $t = \{1, 2, \ldots T\}$ **do**
    $C_t$ = $\epsilon$(D,$D_t$) {Create classifier $C_t$}
    $p_{kt}(x)$ = Prob($y = k|x$), $k = \{1, 2, \ldots K\}$
    $h_{kt}(x) \leftarrow (K$ - 1)($logp_{kt}(x)$ - $\frac{1}{K}$. $\sum_{k'} logp_{k't}$(x))
    $D_i(x) \leftarrow D_i(x)$.exp($\frac{1 - K}{K}$.$y_i^{\mathrm{T}}$.$log(p_t(x_i))$ ) {Update the distribution $D_t$, $i = \{1, 2, \ldots m\}$ }
    Renormalize $D_t(x)$
**end for**
```

plots have been displayed in Figures 3 to 6. The entire list of digital biomarkers along with their corresponding box plots have been included in the supplementary material. Note that the number of digital biomarkers correspond to the number of the box plots which in turn corresponds to number of features selected.

The pipeline developed for this study consists of mul-

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

TABLE 5
Preprocessing techniques

Method	Description
Missing Value Ratio	Removal of data columns where the ratio of missing values is greater than a set threshold
Low Varience Filter	Removal of normalized data columns where the variance is lesser than a set threshold
Highest correlation filter	Removal of data columns which are highly correlated leading to redundancy
Principle Component Analysis	Transformation of data to maximize variance under constraints
Fast Independent Component Analysis	Decomposition of signals to focus on mutual independence of data
Factor Analysis	Generating a common feature by reducing number of common variables

TABLE 6
Algorithms for traditional and ensembled classification and regression

Index	Algorithm Name	Class	Purpose
CT1	Bagged Decision Tree	Traditional	Classification
CT2	Balanced Bagged Decision Tree	Traditional	Classification
CT3	Bagged Random Forest	Traditional	Classification
CT4	Balanced Bagged Random Forest	Traditional	Classification
CT5	Decision Tree	Traditional	Classification
CT6	K-Nearest Neighbours	Traditional	Classification
CT7	Neural Network	Traditional	Classification
CE1	AdaBoost with Decision Tree	Ensemble	Classification (SR)
CE2	AdaBoost with Decision Tree	Ensemble	Classification (S)
CE3	AdaBoost with SVM	Ensemble	Classification (SR)
CE4	AdaBoost with SVM	Ensemble	Classification (S)
CE5	RUSBoost with Decision Tree	Ensemble	Classification (SR)
CE6	RUSBoost with Decision Tree	Ensemble	Classification (S)
CE7	RUSBoost with Random Forest	Ensemble	Classification (SR)
CE8	RUSBoost with Random Forest	Ensemble	Classification (S)
CE9	RUSBoost with SVM	Ensemble	Classification (SR)
CE10	RUSBoost with SVM	Ensemble	Classification (S)

Algorithm 4 : Ensemble of Ensemble: SFORCE

```
// Input: Data Set D = {$(x_1,y_1)$,$(x_2,y_2)$,$\ldots$ ($(x_m,y_m)$}, Feature Set F, Randomization Factor R, Number of trees T,Number of Learning Rounds T′, Learning Algorithm $\epsilon$
// Output: Root node of i$^{\text{th}}$ Boosted Tree
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Random Forest
for $\forall i \in \{1,2,\ldots T\}$ do
    $N_i$ ← Root node of i$^{\text{th}}$ tree
    if All targets belong to same class i.e $y_i$ or $F \in \emptyset$ then
        Call SAMME.R with $N_i$
    end if
    $D^i$ ← bootstraped sample from D
    for Each node do
        $f$ ← Randomly selected $R$ features from $F$
        $N_f$ ← Best Feature from $f$ features
        $N_p$ ← Best Split based on $N_f$
        Call SAMME.R with $N_i$
    end for
end for
 return $N_i$
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
SAMME/SAMME.R
$D_1(x)$ = 1/m {Initialize the weight distribution}
for $t = \{1,2,\ldots T\}$ do
    $C_t$ = $\epsilon$(D,$D_t$) {Create classifier $C_t$}
    $p_{kt}(x)$ = Prob($y = k|x$), $k = \{1,2,\ldots K\}$
    $h_{kt}(x)$ ← $(K$ - 1)($logp_{kt}(x)$ - $\frac{1}{K}$. $\sum_{k'} logp_{k't}$(x))
    $D_i(x)$ ← $D_i(x)$.exp($\frac{1-K}{K}$.$y_i^{\mathrm{T}}$.$log(p_t(x_i))$ ) {$i = \{1,2,\ldots m\}$}
    Renormalize $D_t(x)$
    Call Random Forest with ($\sum_{t=1}^{T'} \alpha_t.C_t$)
end for
```

Algorithm 5 untitled

```
//Input Image dataset $D_{\text{n}}$ and masks $D_{\text{m}}$
//Output Predicted Class
for Each image $i$ in $D_{\text{n}}$ do
    Convert image to a suitable format using conversion software
    Call the pre-processsing techniques on the formatted images
    Using mask $j$ for corresponding $i$, extract radiomics features
    Create a grid of p-values
    for EACH value in grid do
        Call Algorithm 4 with related feature set
    end for
    Obtain accuracy levels and digital bio-markers
end for
```

tidisciplinary stages with involvement of both Radiomics and modern statistics. While Radiomics provides a real-world application based avenue, statistical tools were used to narrow down our biomarker search process. The aim of condensing the number of features is to preserve the features with the highest level information embedded in them.

However it must also be duely noted that this pipeline is quite delicate when it comes to producing results as the errors encountered in each step are rippled onto the next stages. Furthermore an increased sample dataset size could help further fine tune the model. Additional Deep Learning frameworks can also be introduced to provide competition to the incumbent design model.

References

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

TABLE 7
Performance Analysis

Model	CT1	CT2	CT3	CT4	CT5	CT6	CT7	CE1	CE2
Accuracy Reading	0.9917	0.9870	0.9959	0.9959	0.9651	0.9949	1.0000	0.8713	0.9709
Time Taken	44.2017	44.2017	27.9943	27.9943	15.5171	18.6339	26.5288	127.2628	127.2628
Model	CE3	CE4	CE5	CE6	CE7	CE8	CE9	CE10	**SFORCE (SR) with K-Fold cross validation**
Accuracy Reading	1.0000	1.0000	0.9870	0.9896	1.0000	0.9977	0.9920	0.9977	**0.9974**
Time Taken	54.6694	54.6694	156.7184	156.7184	24.2733	24.2733	27.1538	27.1538	**570.5684**

TABLE 8
TN vs Non-TN

P-Value	Number of features	SAMME Accuracy	SAMME.R Accuracy
1	20	81.25	90.39
0.5	16	90.39	93.25
0.1	6	75	81.25

TABLE 9
HER vs Luminal-A vs TN

P-Value	Number of features	SAMME Accuracy	SAMME.R Accuracy
1E-5	16	72	63.63
1E-6	15	70	72.7
17-5	13	72.7	70

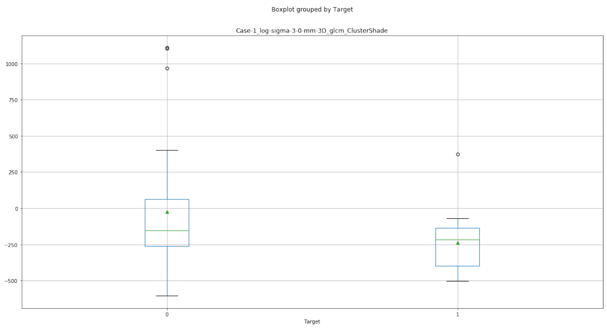

Fig. 3.

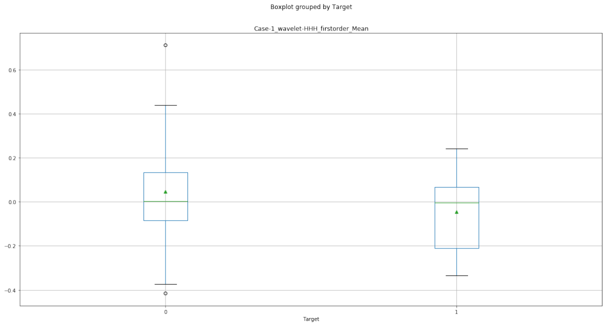

Fig. 4.

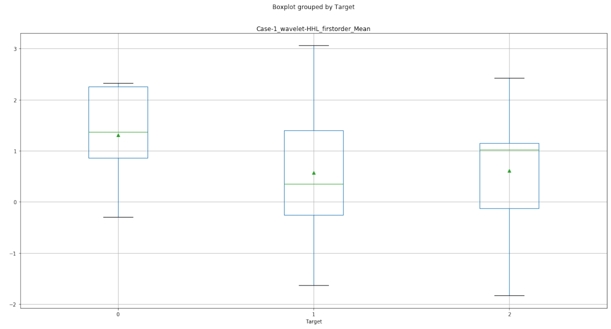

Fig. 5.

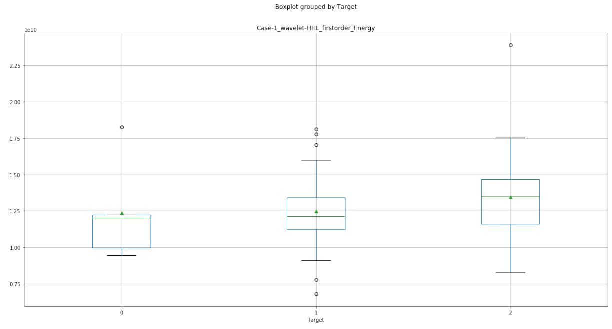

Fig. 6.