DigiOnco: A Pipeline to Unveil Digital Non-Invasive Biomarkers from Multi-parametric Radiomics Footprints for the Indian Breast Cancer Cohort with Multiple Molecular Subtypes

Santhi Natarajan, Anand Ravishankar, Bharathi Malakreddy A[a], G.Lohith, Kritika Sekar, Shivakumar Swamy, Kumar Kallur, Basavalinga Ajai Kumar, Mahesh Bandimegal, Krithika Murugan[b]

[a]*RadioGenomics Research Group, BMS Institute of Technology and Management, Visweswaraiah Technological Univesity, Bangalore, India*
[b]*Health Care Global (HCG) Hospitals, Bangalore, India*

Abstract

Background: Digital Imaging and Analytical Models (DIAMs) assisted Radiomics has emerged as a promising tool towards offering personalized therapeutics to patients. Oncological researchers seem to have benefited the most from the automated and reproducible analytics that DIAMs offer. However, statistical deductions and associations of inferences from DIAMs have to be cross-validated with a higher rate of concordance and robustness. With the amount of digital healthcare data amounting to 2.5 Zetabytes as on date, there is a dire need for standardization of DIAMs applied to cohort based study with heterogeneous patient centric parameters, including age, phenotypes and genotypes.

Methods: DigiOnco, is a novel pipeline, intricately woven with carefully chosen set of Machine Learning (ML) algorithms. DigiOnco unveils the digital non-invasive biomarkers from multi-parametric 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. An independent cohort, having a set of external group of patients identified with breast cancer across multiple molecular subtypes, is used for the study.

Findings: The DigiOnco pipeline consists of multi-disciplinary stages involving both Radiomics and modern statistics. While Radiomics provides a real-world application based approach, statistical tools help to narrow down our biomarker search process. DigiOnco offers accuracy levels ranging from 72.7% to 93.25%, in mapping the feature sets to the various molecular subtypes of breast carcinoma. The internal cross validation and the retrospectively validated study on the cohort affirm our hypothesis.

Interpretation: Considering the integration of our current findings with follow-up studies branching into other medical sub-doamins, the potential of homogenizing Machine Learning is huge in this field. The aim of obtaining the subset of optimal features from the radiomics feature set extracted from imaging, is to preserve the features with the highest level of information embedded in them. However it must also be duly noted that this pipeline is quite delicate when it comes to producing results, as the errors encountered in each step are rippled onto the subsequent stages. This is mitigated by having an optimal p-value level for the corresponding feature set. Furthermore, an increased sample size could help further fine tune the model. Additional Deep Learning frameworks can also be introduced to provide competition to the incumbent design model.

Funding: Vision Group of Science and Technology, Government of Karnataka, India

Keywords: Non-invasive biomarkers, Radiomics, PET-CT imaging, Machine Learning, Oncology

1. Introduction

1.1. Personalized Therapeutics: An oveview

Personalized medicine, diagnostics and therapeutics uphold the promise of accurate decision making by leveraging the power of machine learning/deep learning based Digitized Imaging and Analytical Models (DIAM) [1][2][3][4]. 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, 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 [5].

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, tumor heterogenity has made therapeutic decision making more complex a process. Molecular sub-typing and classification of breast carcinoma now involves multi-modal analytics, including characterizing the morphology and grading a tumor, histopathology, Immunohistochemistry (IHC) and in situ hybridization (ISH) [6]. 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 [7].

1.3. Radiomics Assisted Digitization: An aid for Oncological Decision Making amidst Heterogeneity

With the wide-spread know how of Computed Tomography (CT), Positron Emission Tomography (PET), and Magnetic Resonance Imaging (MRI) methods, Radiomics gathered much attention in the last few years of oncological research [8]. 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) [9] 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 [10].

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

The stability and reproducibility of the existing Radiomics models [11] [12], 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 1 are predominantly applied to MRI imaging, due to its monochromatic image quality and wide availability of literature in terms of statistical image analytics.

Table 1: Previous Works

Previous Work	**Result Description**
Xie T, et al. [9]	Using machine learning analysis of multiparametric MR radiomics to classify immunohistochemical (IHC) subtypes of breast cancer
Liu YX, et al. [7]	Minimal use of IHC markers to distinguish three different subtypes of breast cancer displaying diverse prognostic characteristics
Robertson S, et al. [3]	Use of AI and Deep Learning in diagnostic breast pathology and other recent digital image analysis
Heather D Couture, et al. [1]	Using Deep Learning to predict complex properties such as ER status, histologic subtype, and intrinsic subtype
Jaber MI, et al. [4]	Identification of cancer-rich patches among multiscale patches in H&E-stained WSIs that can be generalized to any indication.

The predominant issue in reproducing and comparing results across multiple studies stems from the challenge of having enriched data. Having a common data gathering point can provide an extensive and accurate comparison of different studies conducted till date. This need-of-the-hour problem needs to be dealt with a collaborative approach. As per a 2020 Stanford study, the amount of healthcare data including the Radiomics and Radiogenomics analytics data is annually growing at a steady rate of 48% which indicates the opportunity for collaborations [13].

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 Indian breast cancer cohort, having a set of external group of patients.

2. Methods

DigiOnco based imaging analytics in cancer care holds immense potential to unlock valuable clinical insights from an abundance of patient imaging records, aided by sophisticated modeling, which will lead to deeper personalization of cancer treatment and in-turn improved outcomes. Presently, the oncologist is overwhelmed with scientific literature, swiftly evolving treatment techniques and the exponentially increasing amount of clinical data. With the huge deluge in imaging data, it becomes increasingly difficult to translate this data into structured information. DigiOnco processes imaging data and extracts features in the form of digital biomarkers. These, along with the clinically derived biomarkers (blood, tissue, imaging etc), guides the oncologist

in clinical decision-making during routine clinical practice to more accurately predict and prognosticate cancer treatment strategies.

Breast cancer is high in intratumoral heterogeneity. The different molecular subtypes of breast cancer respond differently to surgery, radiation or chemo-hormonal treatments [14]. Therefore, recognizing imaging markers directly to distinguish molecular subtypes, without invasive biopsies, would help in guiding treatment plans for breast cancers. DigiOnco effectively improves sensitivity and specificity of breast cancer diagnosis from PET-CT imaging phenotypes, with the help of quantitative imaging characterization. This method can support comprehensive evaluation of heterogeneity of the lesions and predict the prognosis in advance. DigiOnco based radiomic analysis of biological characteristics can effectively differentiate different subtypes of breast cancer with good accuracy and this approach serves as a more convenient and non-invasive biomarker for the prediction of breast cancer subtypes.

2.1. Obtaining Raw data

In order to obtain distinct yet comparable subjects, an Indian breast cancer cohort with multiple molecular subtypes of 89 patients was selected for this study. The dataset consists of four intrinsic molecular subtypes of breast cancer which are contrasted on the genes a cancerous cell expresses [15] [16]. The dataset has been descibed in Table 2.

For every patient, a CT scan was conducted to obtain cross-sectional images of the hypothesised tumor location. PET-CT scans provide a more detailed description of the patient's condition by increasing the radiation level the patient is exposed to. We obtained the DICOM (Digital Imaging and Communication in Medicine) images from the scan in the three views namely, Axial, Sagittal and Coronal. For each patient, 323 new studies were conducted with each study having 384 series which corresponded to 466 instances or images of the scan.

2.2. Convert to a suitable format

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 in a more compact scheme [17]. 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 [18], which is an open source software for image computation. Plastimatch performs rasterization by processing the DICOM image which is described in a polyline vectorized format, and converting it into a series of pixels. The subroutine for rasterization of a DICOM image set with coordinates 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
```

Table 2: 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]

2.3. *Obtaining Radiomics Features*

Directly extracting information from images has certain drawbacks. Consider tumor classification using a standard Convolutional Neural Network (CNN). The CNN might be extremely successful in determining the existence of a blob of mass and its exact location. However, diagnosing the precise nature and feature set of the tumor is extremely difficult for a CNN as it views the image as a collection of pixels without any regard for the information embedded in the image.

To tackle this issue, we have utilized radiomics based libraries and algorithms to extract feature sets from the medical images to reveal hidden characteristics. The open-source Python library, PyRadiomics [19], 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 dataset to provide a comprehensive view of the data. The filters applied are listed in table 3. The processing and subsequent image

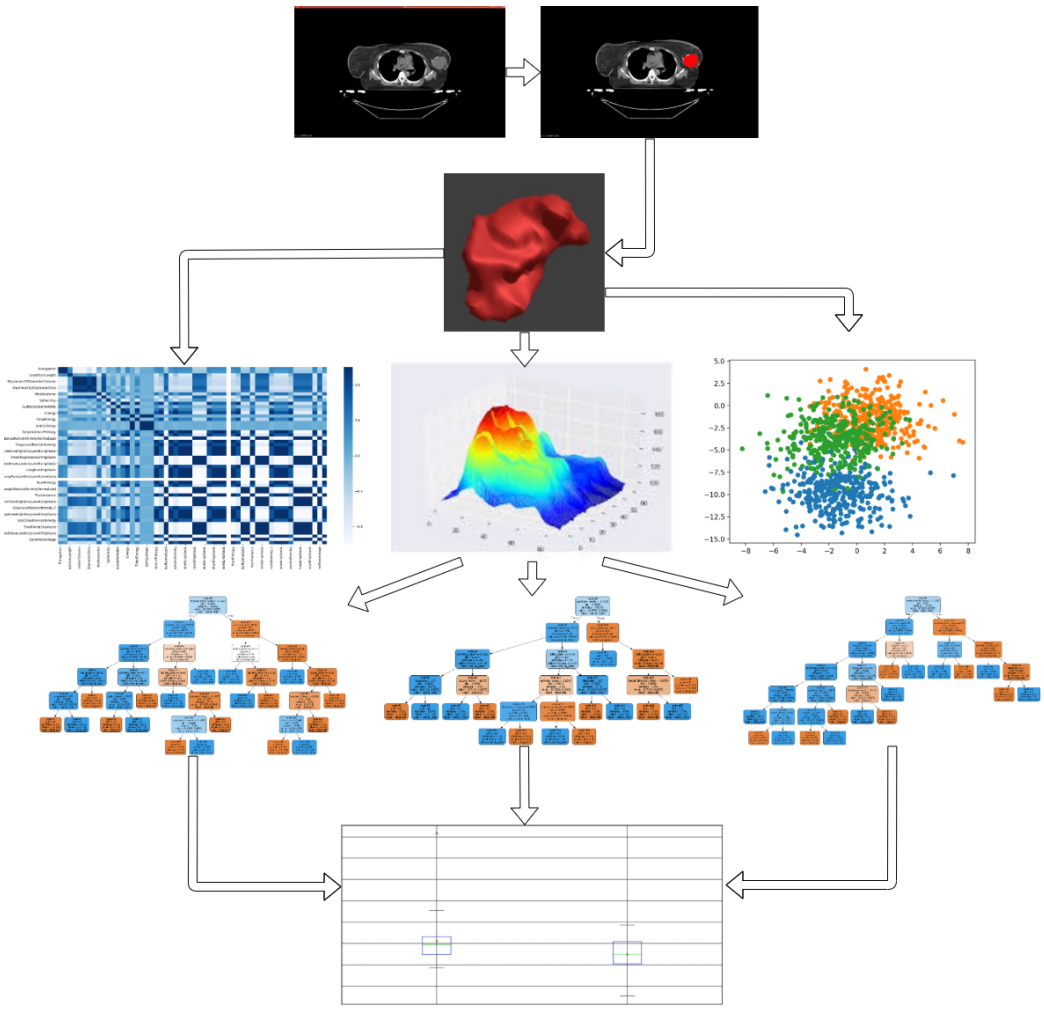


Figure 1: Data-Flow

characteristics are discussed in the Supplementary Material.

PyRadiomics extracts radiomics features from the CT scan 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 3 are applied using SimpleITK [20], PyWavelets [21], and Numpy [22]. 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 constraints 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) [23] and have have been shown in Tables 4 and 5.

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. Once the entire feature set has been collected, the classification task can begin.

2.4. Applying Pre-processing Techniques

From the 1700 features collected, we select a sub-set of top ranking features which contribute significantly to the classification task. 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. We have applied multiple data pre-processing techniques to the feature set, which are listed in Table 6.

Since the number of test subjects for each class is different, a threshold confidence level must be specified during the hypothesis testing phase. A 'P-value' is used to evaluate the hypothesis under observation. A lower p-value corresponds to a higher confidence level in the predictions, thereby accommodating only the information-rich features. The number of features selected after the pre-processing step is directly proportional to the p-value. We created a grid for varying p-values and derived the corresponding number of features for each p-value.

2.5. Model-based Predictions

Once the features have been narrowed down, we initiated the model building process. For any task in 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 7. In order to determine which algorithm would perform the best for our cohort dataset, we trained all the

Table 3: Applied Filters

Filter	**Description**	**Equation**
Wavelet	Selective emphasizing de-emphasizing of image	-
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 fliter for a σ value	$\frac{1}{(\sigma\sqrt{2\pi})^3} e^{-\frac{x^2+y^2+z^2}{2\sigma^2}}$
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	-

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 8.

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 simultaneously, aiding each other in the prediction process. Random Forests provides a strong learning system with the occasional pitfall of overfitting. The algorithms are discussed in-depth in the supplementary material.

To obtain digital bio-markers, we conducted two case studies on 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 its char-

Table 4: Features-I

Feature Class	**Feature**	**Feature Class**	**Feature**	**Feature Class**	**Feature**
Shape	Max_2D_Diameter_C Max_2D_Diameter_R Max_2D_Diameter_S 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	Grey Level Size Zone Matrix	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 Non_Uniformity Non_Uniformity_Normalized Variance
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 Normalized_Uniformity Run_Percentage Run_Variance Short_Run_Emphasis Short_Run_High_Emphasis Short_Run_Low_Emphasis Uniformity	Gray Level Size Zone Matrix	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 Dependence_Entropy Dependence_Non_Uniformity
Neighbouring Gray Tone Difference Matrix	Busyness Coarseness Complexity Constrast Strength				

acteristics with those of Luminal A. We trained the model to place the test subjects into the Luminal-A class, resulting in an accuracy of 72.7%. The results for different p-values are described in Tables 9 and 10. Based on these results, we obtained the box-plots for the selected features which act as bio-markers for future reference.

2.6. Role of source funding

The funders had no role in study design, data interpretation, writing of the manuscript, and decision to submit. All the authors had full access to all the data used in the study and had final

Table 5: 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 6: Preprocessing techniques

Method	**Description**
Missing Value Ratio	Removal of data columns where the number of missing values $\geq$ threshold
Low Varience Filter	Removal of normalized data columns where the variance $\leq$ threshold
Highest correlation filter	Removal of data columns which are highly correlated leading to redundancy
Principle Component Analysis	Transformation of data to maximize σ^2 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 7: 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)

Table 8: 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**
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**

responsibility for the decision to submit for publication.

3. Results and Conclusion

From the data-driven pipeline, we obtained the quantifiable digital biomarkers in the form of box and whisker plots. These plots provide a convenient method of displaying the data distribution and provide insight to the oncological expert during prognosis of future test subjects. Sample box plots are displayed in Figures 2 to 5. The entire list of digital biomarkers along with their corresponding box plots are 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 mappings between features and filters selected.

Table 9: TN vs Non-TN

P-Value	**Number of** Features	**Accuracy** (SAMME)	**Accuracy** (SAMME.R)
1	20	81.25	90.39
0.5	16	90.39	93.25
0.1	6	75	81.25

Table 10: HER vs Luminal-A vs TN

P-Value	**Number of** Features	**Accuracy** (SAMME)	**Accuracy** (SAMME.R)
1E-5	16	72	63.63
1E-6	15	70	72.7
1E-7	13	72.7	70

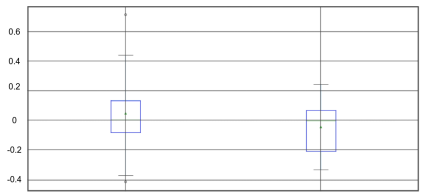


Figure 2: Sample Box and wiskhers plot for TN (left) vs Non-TN (right)

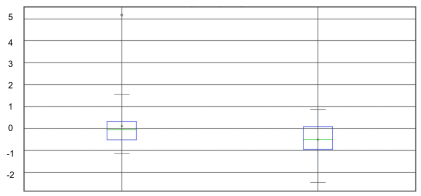


Figure 3: Sample Box and wiskhers plot for TN (left) vs Non-TN (right)

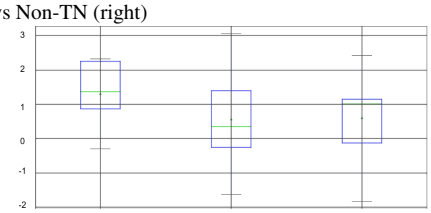


Figure 4: Sample Box and wiskhers plot for HER (left) vs Luminal-A (center) vs TN (right)

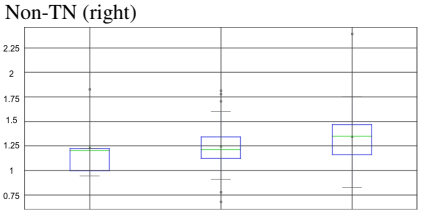
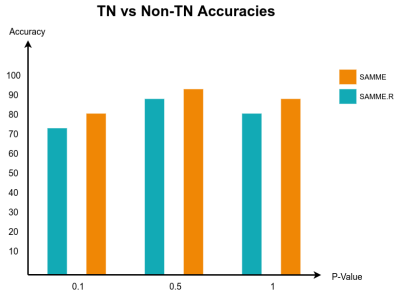
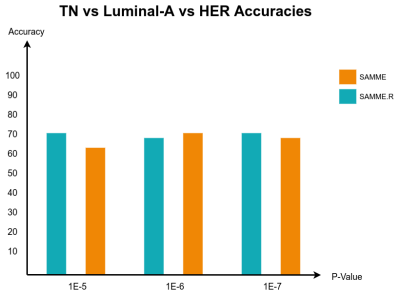


Figure 5: Sample Box and wiskhers plot for HER (left) vs Luminal-A (center) vs TN (right)





The pipeline developed for this study consists of multidisciplinary stages with involvement of both Radiomics and modern statistics [24]. 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.

3.1. Contributors

The Radiogenomics Research Group of BMS Institute of Technology and Management developed the Image Processing Pipeline with Machine Learning algorithmic steps, and has also executed them on computing platforms, producing the results discussed in the paper. The oncological research team within the HCG group served as the medical partner and provided the data for Indian breast cancer cohort with multiple molecular subtypes, as well as other ethical clearances for conducting this cross-reference study on the cohort. All authors have contributed to the formulation of the reseaarch statement, methodology, installation, execution and verfication

of the findings and mapping of results to inferences. All authors contributed to the writing of the final version of the Article.

3.2. Declaration of interests

We declare no competing interests.

3.3. Data Sharing

3.4. Acknowledgment

The authors would like to thank the Vision Group on Science and Technology, Government of Karnataka, India, for funding this project under the Centre for Design and Research on Healthcare Applications using AI. The authors also would like to thank the HCG management for support provided to obtain the ethical clearances for this study.

References

[1] Heather D. Couture1 Lindsay A. Williams, Joseph Geradts Sarah J. Nyante Ebonee N. Butler J. S. Marron, Charles M., Perou Melissa A. Troester and Marc Niethammer. Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. Nature Partner Journals (2018)

[2] Lohmann, P., Bousabarah, K., Hoevels, M. et al. Radiomics in radiation oncology—basics, methods, and limitations. Strahlenther Onkol 196, 848–855 (2020). https://doi.org/10.1007/s00066-020-01663-3.

[3] Robertson S, Azizpour H, Smith K, Hartman J. Digital image analysis in breast pathology-from image processing techniques to artificial intelligence. Transl Res. 2018 Apr;194:19-35. doi: 10.1016/j.trsl.2017.10.010. Epub 2017 Nov 7. PMID: 29175265.

[4] Jaber MI, Song B, Taylor C, Vaske CJ, Benz SC, Rabizadeh S, Soon-Shiong P, Szeto CW. A deep learning image-based intrinsic molecular subtype classifier of breast tumors reveals tumor heterogeneity that may affect survival. Breast Cancer Res. 2020 Jan 28;22(1):12. doi: 10.1186/s13058-020-1248-3

[5] Carelli S, Giallongo T, Rey F, Barzaghini B, Zandrini T, Pulcinelli A, Nardomarino R, Cerullo G, Osellame R, Cereda C, Zuccotti GV, Raimondi MT. Neural precursors cells expanded in a 3D micro-engineered niche present enhanced therapeutic efficacy in vivo. Nanotheranostics 2021; 5(1):8-26. doi:10.7150/ntno.50633. Available from https://www.ntno.org/v05p0008.htm

[6] Blows, Fiona & Driver, Kristy & Schmidt, Marjanka & Broeks, Annegien & Leeuwen, Flora & Wesseling, Jelle & Cheang, Maggie & Gelmon, Karen & Nielsen, Torsten & Blomqvist, Carl & Heikkilä, Päivi & Heikkinen, Tuomas & Nevanlinna, Heli & Akslen, Lars & Bégin, Louis & Foulkes, William & Couch, Fergus & Wang, Xianshu & Cafourek, Vicky & Huntsman, David. (2010). Subtyping of Breast Cancer by Immunohistochemistry to Investigate a Relationship between Subtype and Short and Long Term Survival: A Collaborative Analysis of Data for 10,159 Cases from 12 Studies. PLoS medicine. 7. e1000279. 10.1371/journal.pmed.1000279.

[7] Liu YX, Wang KR, Xing H, Zhai XJ, Wang LP, Wang W. Attempt towards a novel classification of triple-negative breast cancer using immunohistochemical markers. Oncol Lett. 2016 Aug;12(2):1240-1256. doi: 10.3892/ol.2016.4778. Epub 2016 Jun 23. PMID: 27446423; PMCID: PMC4950427.

[8] Arnaud Marcoux, Ninon Burgos, Anne Bertrand, Marc Teichmann, Alexandre Routier, et al.. An Au- tomated Pipeline for the Analysis of PET Data on the Cortical Surface. Frontiers in Neuroinformatics, Frontiers, 2018, 12, ff10.3389/fninf.2018.00094ff. ffhal-01950933ff

[9] Xie T, Wang Z, Zhao Q, et al. Machine Learning-Based Analysis of MR Multiparametric Radiomics for the Subtype Classification of Breast Cancer. Front Oncol. 2019;9:505. Published 2019 Jun 14. doi:10.3389/fonc.2019.00505

[10] Shaikh FA, Kolowitz BJ, Awan O, Aerts HJ, von Reden A, Halabi S, Mohiuddin SA, Malik S, Shrestha RB, Deible C. Technical Challenges in the Clinical Application of Radiomics. JCO Clin Cancer Inform. 2017 Nov;1:1-8. doi: 10.1200/CCI.17.00004. PMID: 30657374.

[11] van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging-"how-to" guide and critical reflection. Insights Imaging. 2020;11(1):91. Published 2020 Aug 12. doi:10.1186/s13244-020-00887-2

[12] Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2016 Feb;278(2):563-77. doi: 10.1148/radiol.2015151169. Epub 2015 Nov 18. PMID: 26579733; PMCID: PMC4734157.

[13] Panayides, Andreas & Pattichis, Marios & Leandrou, Stephanos & Pitris, Costas & Constantinidou, Anastasia & Pattichis, C.. (2019). Radiogenomics for Precision Medicine With A Big Data Analytics Perspective. IEEE Journal of Biomedical and Health Informatics. 23. 2063-2079. 10.1109/JBHI.2018.2879381.

[14] Tsougos, Ioannis & Vamvakas, Alexandros & Kappas, Constantin & Fezoulidis, Ioannis & Vassiou, Katerina. (2018). Application of Radiomics and Decision Support Systems for Breast MR Differential Diagnosis. Computational and Mathematical Methods in Medicine. 2018. 1-8. 10.1155/2018/7417126.

[15] Sengal AT, Haj-Mukhtar NS, Elhaj AM, Bedri S, Kantelhardt EJ, Mohamedani AA. Immunohistochemistry defined subtypes of breast cancer in 678 Sudanese and Eritrean women; hospitals based case series. BMC Cancer. 2017 Dec 1;17(1):804. doi: 10.1186/s12885-017-3805-4. PMID: 29191181; PMCID: PMC5710067.

[16] Tang P, Tse GM. Immunohistochemical Surrogates for Molecular Classification of Breast Carcinoma: A 2015 Update. Arch Pathol Lab Med. 2016 Aug;140(8):806-14. doi: 10.5858/arpa.2015-0133-RA. PMID: 27472239.

[17] Larobina M, Murino L. Medical image file formats. J Digit Imaging. 2014;27(2):200-206. doi:10.1007/s10278-013-9657-9

[18] Sharp, G. & LI, R. & Wolfgang, John & Chen, G. & Peroni, Marta & Spadea, Maria & Mori, Shinichiro & Zhang, J. & Shackleford, J. & Kandasamy, Nagarajan. (2010). PLASTIMATCH– AN OPEN SOURCE SOFTWARE SUITE FOR RADIOTHERAPY IMAGE PROCESSING.

[19] van Griethuysen, J. J. M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G. H., Fillon-Robin, J. C., Pieper, S., Aerts, H. J. W. L. (2017). Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Research, 77(21), e104–e107. 'https://doi.org/10.1158/0008-5472.CAN-17-0339 ¡https://doi.org/10.1158/0008-5472.CAN-17-0339¿

[20] R. Beare, B. C. Lowekamp, Z. Yaniv, "Image Segmentation, Registration and Characterization in R with SimpleITK", J Stat Softw, 86(8), doi: 10.18637/jss.v086.i08, 2018.

[21] Gregory R. Lee, Ralf Gommers, Filip Wasilewski, Kai Wohlfahrt, Aaron O'Leary (2019). PyWavelets: A Python package for wavelet analysis.

[22] Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357–362 (2020).

[23] Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, Ashrafinia S, Bakas S, Beukinga RJ, Boellaard R, Bogowicz M, Boldrini L, Buvat I, Cook GJR, Davatzikos C, Depeursinge A, Desseroit MC, Dinapoli N, Dinh CV, Echegaray S, El Naqa I, Fedorov AY, Gatta R, Gillies RJ, Goh V, Götz M, Guckenberger M, Ha SM, Hatt M, Isensee F, Lambin P, Leger S, Leijenaar RTH, Lenkowicz J, Lippert F, Losnegård A, Maier-Hein KH, Morin O, Müller H, Napel S, Nioche C, Orlhac F, Pati S, Pfaehler EAG, Rahmim A, Rao AUK, Scherer J, Siddique MM, Sijtsema NM, Socarras Fernandez J, Spezi E, Steenbakkers RJHM, Tanadini-Lang S, Thorwarth D, Troost EGC, Upadhaya T, Valentini V, van Dijk LV, van Griethuysen J, van Velden FHP, Whybra P, Richter C, Löck S. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology. 2020 May;295(2):328-338. doi: 10.1148/radiol.2020191145. Epub 2020 Mar 10. PMID: 32154773; PMCID: PMC7193906.

[24] Giraud P, Gasnier A, et al. Radiomics and Machine Learning for Radiotherapy in Head and Neck Cancers. Front Oncol. 2019;9:174. Published 2019 Mar 27. doi:10.3389/fonc.2019.00174