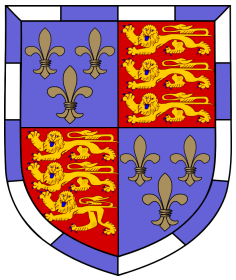
UNIVERSITY OF
CAMBRIDGE

Feasibility and sensitivity study of radiomic features in photoacoustic imaging of patient-derived xenografts

Ivo Vladislavov Petrov



St. John's College
June 30, 2024

Word count: 6813 (using texcount)

Contents

List of Figures

Chapter 1

Introduction

Cancer is a serious disease where abnormal cell growth can cause significant health issues and greatly affect the lives of patients. In particular, breast cancer is among the most widespread forms of tumours, affecting around 2.3 million people every year [2], making it the second most common. This study focuses on the luminal B and basal types, which represent around 30-40% of all breast cancer cases. The classification is based on 3 hormone receptors: estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor (HER2) [3, 4]. The luminal B class is positive for ER but low or negative PR and HER2, while the basal one is negative across all three. The 5-year survival rate varies significantly, with luminal B at $\approx 91\%$ survivability, in contrast to the 77.1% of the basal subtype [3], highlighting the need for accurate differentiation for an effective treatment and accurate prognosis.

An emerging technology for non-invasive imaging of tissue is the use of photoacoustic imaging (PAI) [5, 6]. Previous studies have shown this technology's potential for being able to highlight canncerous regions [7, 8, 9], and being non-ionizing in nature, it presents a safe and informative diagnostic methodology. While PAI images are informative for inspection by a pathologists, in-depth analysis is required for determining the subtype of cancer. As a way to capture structural information about the tissue, a set of features known as radiomics [10, 11, 12] can be extracted, which have been paired with other modalities in clinical settings [13, 14, 15]. These features offer crucial insights into the scans that cannot be perceived visually. However, for the purpose of PAI, their applicability in in-vivo samples has not been sufficiently explored. Such technology can be useful for improving the quality and efficiency of the prediction and prognosis. Escudero Sanchez et al. [16] proposed a framework to determine the feasibility of radiomic features in a setting including the aforementioned tumour models. In this work we aim to replicate the results of their research, and evaluate its consistency across different pipeline modifications.

Chapter 2

Background

In this chapter, we briefly introduce concepts that are important to understand the nature of the data and our work.

2.1 Related work

2.1.1 Photoacoustic Imaging

Photoacoustic (or optoacoustic) imaging, which we will henceforth refer to as **PAI**, has recently seen major development, both in a technical and engineering aspect, as well as on an application level [17]. PAI is based on the photoacoustic effect [18], through which molecules (and hence, tissue) can transform optical energy into audible pressure waves that propagate through the body. Usually, electromagnetic radiation in the visible or near-infrared is used, as it has sufficient penetrative capabilities, and is also non-ionizing [17, 5]. Such frequencies are useful to detect the absorption, and hence the concentration, of hemoglobin, oxidized hemoglobin, lipids and water [17], revealing different structures in the tissue. The resulting oscillations from the absorption produce sound waves that are detected from an array of transducers [19]. Paired with computed tomographical methods, the signal is used to produce a three-dimensional image of the underlying tissue, with vertical and lateral resolutions in the order of 50 μm, and a depth penetration of $> 1cm$ [5].

2.1.2 Radiomic Features

The study of radiomic features [12] is an emerging field that aims to quantifiably enhance medical imaging through features that summarize properties of the underlying imaging data [11, 20]. Prior research has shown promising applications of radiomic features, most prominently in oncological settings. The radiomics aim to capture discrepancies in the signal distributions that cannot be perceived by the human eye, but can be efficiently combined with analytical AI methods [21, 22]. Radiomic features on photoacoustic imaging have previously been seldom explored, meaning the framework described in this work is an attempt to extend the previous research done in Escudero Sanchez et al. [16]. In particular, we examine *first-order histogram statistics* [20], *shape features* [23], *Gray Level Co-occurence Matrix* (GLCM) [24], *Gray Level Run Length Matrix* (GLRLM) [25], *Gray Level Size Zone Matrix* (GLSZM) [26], *Neighbouring Gray Tone Difference Matrix* (NGTDM) [27] and *Gray Level Dependence Matrix* (GLDM) [28] features. All of them are implemented in the Python package `pyradiomics` [29] , which was used to generate the relevant dataset. Below we describe the general approach each of the feature

categories covers, with a comprehensive list of each feature found in the `pyradiomics` package documentation [29], or in the supplementary materials in the original work.

2.2 Materials

In this section we describe the methodology with respect to gathering and processing the data up to retrieving the radiomic features.

Animal and tumour treatment A total of 21 cases of breast cancer were investigated, with 10 of them being of the basal patient-derived xenograft (PDX) model, and 11 – of the luminal B PDX model. We feature no control group, as the focus of this work is to evaluate the discriminative power between untreated tumours of the aforementioned models. The tissue fragments were provided by the Caldas laboratory at the Cancer Research UK Cambridge Institute [16] in a cryogenically frozen state. They were defrosted at 37° washed with Dulbecco's modified eagle's medium (41965039, Gibco) [16], mixed with matrigel (354262, Corning®, NY, USA) [16] and surgically implanted into 6-9 week-old female NSG mice (Jax Stock 005557) [16] as per standard protocol [16, 30]. The animals were kept in a sterile enviroment in hermetically sealed cages with individual filtered air supply, in half-day light cycles, with the appropriate amount of food and water. The tumours were monitored with callipers measuring the mean diameter, until it reaches a length of ≈1 cm. At that point, photoacoustic imaging was performed on the mice, which were thereafter euthanised.

Information regarding licensing, ethical approval and further guidelines can be found in the App. A.2.

Photoacoustic imaging Multispectral Optoacoustic Tomography (MSOT) was used to perform PAI, using the inVision 256-TF scanner [31], a diagram and image of which can be seen in App. A.1, as provided by Tong et al. [1]. The scanner utilizes a laser that can emit pulses at a wavelength of anywhere between 660 and 1300 nm through 5 optic fibre bundles to achieve uniform illumination. The transducer array covers an angle of 270°, with the resulting signal being reconstructed using CT-based techniques. Images were acquired at a depth of $\approx 3cm$, achieving a lateral resolution of $\approx 190\mu m$.

Prior to the imaging process, each mouse was anesthesized using 3-5% isoflurane[16] and were shaved to prevent hair-related imaging artifacts. Finally, to ensure uniform and accurate sound wave propagation, ultrasound gel was applied to each subject, prior to wrapping them in a polyethylene membrane. They were further placed in the MSOT system, surrounded by a 36° constant temperature water container.

The device was left to stabilize for 15 minutes in order for the mice's breathing to adapt to the 100% oxygenated air, as well as to allow the medium to stop oscillating. Slices were acquired in an interval of 1 mm, each of which takes approximately 12 seconds. Each slice was imaged using an average of 6 consecutive pulses over 15 wavelengths between 700 and 880 nm. This range covers the isosbestic points of oxyhaemoglobin and deoxyhaemoglobin, allowing the use of their concentration as a biomarker[32, 33].

Image reconstruction was then performed using two different algorithms - backprojection [34], and a model-based algorithm using the ViewMSOT software [35], in order to verify robustness with respect to the reconstruction factor. The former is the most common method for tomographic reconstruction, and is comprised of solving a system of equations describing the wave propagations in Fourier space. On the other hand, the latter achieves reconstruction

through a model-based iterative inversion, where the solution of the wave equation is discretized in the time domain [35]. Producing the images themselves are part of the original work [16], with the derived radiomic features being found in the supplementary repository[1].

Segmentation Segmentation was performed using a combination of MATLAB preprocessing and manual contour drawing around the tumours on each slice, which was performed by an experienced member of the team that produced the original paper, as described by Escudero Sanchez et al. [16].The delineations were determined on the image produced by the backprojection reconstruction with a wavelength of 800 nm, which is an equilibrium point of oxygenated and deoxygenated haemoglobin. The same Volume of Interest (VOI) was then used to extract radiomic features for the remaining wavelengths and reconstruction methods.

Extracting the radiomic features Radiomic feature extraction was performed in an anonimized manner independently of the segmentation delineation, using a quantisation of anywhere between 8 and 256 gray levels. All shapes features were disregarded, as we use the same contours for each sample across one patient case. We do, however, take into account the VOI by eliminating it as a factor, a step in our pipeline that is further elaborated on in Sec. 3.2. A total of 93 volume-based features were computed in the categories mentioned in Sec. 2.1.2:

- **First-order statistics (FOS)**– 18 features
- **Gray Level Co-occurrence Matrix (GLCM)** – 24 features
- **Gray Level Dependence Matrix (GLDM)** – 14 features
- **Gray Level Run Length Matrix (GLRLM)** – 16 features
- **Gray Level Size Zone Matrix (GLSZM)** – 16 features
- **Neighbouring Gray Tone Difference Matrix (NGTDM)** – 5 features

[1]The repository for the original work can be found in the corresponding GitHub page.

Chapter 3

Methodology and Technical Aspects

In this chapter, we describe the technical considerations of our approach to reproducing the results of Escudero Sanchez et al. [16]. We highlight the importance of good data-analytical practices and further extend the work by validating the results under different settings.

3.1 Sensitivity analysis

We first perform a proof-of-concept analysis into the features' sensitivity to different factors. In particular, we perform an analysis of variance (ANOVA) with the Tumour model, reconstruction method, wavelength and the number of grey levels as factors. We consider both an unbalanced analysis, as well as a balanced one by removing a randomly selected luminal specimen, as described in Escudero Sanchez et al. [16]. We measure the fraction of explained variance of feature $\boldsymbol{x}$ under a set factors $\mathcal{F}$, which we denote as $\eta^2_{\mathcal{F}}$, where $\mathcal{F} \subset \mathbb{F} = \{\text{Model}, \text{Recon}, \text{Wavelength}, \text{GL Bins}\}$. For a given subset of factors $\mathcal{F}$, we denote the set of values the factors can take as $\mathbb{P}_{\mathcal{F}}$, the features which have corresponding factor values $v \in \mathbb{P}_{\mathcal{F}}$ as $\boldsymbol{x}_v$, and their number as n_v. Thus, we calculate $\eta^2_{\mathcal{F}}$ as follows:

$$\eta^2_{\mathcal{F}} = \sum_{v \in \mathbb{P}_{\mathcal{F}}} n_v (\bar{\boldsymbol{x}}_v - \bar{\boldsymbol{x}})^2 \tag{3.1}$$

Given this definition, we combine all second-order interactions into an 'Error' term, which we further show do be negligible compared to the first-order factors – $\eta^2_{\text{Error}} = \sum_{\mathcal{F} \subset \mathbb{F}, |\mathcal{F}|>1} \eta^2_{\mathcal{F}}$ Finally, we compute the sensitivity fraction for all first-order factors – $\frac{\eta^2_f}{\sum_{f` \in \mathbb{F} \cup \{\text{Error}\}} \eta^2_{f`}}$ and report this measure for all features. For the factors that are found to be most significant, we repeat the process for fixed values of said factors.

We further perform this analysis in a 5-fold cross-validation setting, removing 3 luminal and 2 basal specimen in each iteration. We reran the η^2 computation for each fold and quote the mean and standard deviation of the metric, as well as the coefficient of variance (CoV) as a measure of sampling robustness.

3.2 VOI Normalization

Before proceeding to investigating the discriminatory potential of each feature with respect to the model, we aim to eliminate any relationship they exhibit with the Volume of Interest (VOI). This step is crucial, as the fact we only have information from 21 patients allows any classifier to

extract the VOI information from the features which is directly related to the underlying model. Allowing this information to pass through may cause any results to be biased towards features with higher VOI correlations, instead of reflecting any inherent properties of the features.

To this end, we make use of the methodology described in Escudero Sanchez et al. [36]. In essence, for any feature $\boldsymbol{x}$, we attempt to find a function $\boldsymbol{f}$, such that $\boldsymbol{g}(\boldsymbol{v}) = \beta_1 \boldsymbol{f}(\boldsymbol{v}) + (\beta_0 | \dots | \beta_0)$ best fits $\boldsymbol{x}$, where $\boldsymbol{v}$ is the corresponding number of voxels for each feature instance and β_1 and β_0 are free parameters. We consider the following instances of $\boldsymbol{f}$: $\boldsymbol{f}(\boldsymbol{v}) = \boldsymbol{v}$, $\boldsymbol{f}(\boldsymbol{v}) = \boldsymbol{v}^2$, $\boldsymbol{f}(\boldsymbol{v}) = \boldsymbol{v}^3$, $\boldsymbol{f}(\boldsymbol{v}) = \boldsymbol{v}^{-1}$, $\boldsymbol{f}(\boldsymbol{v}) = \boldsymbol{v}^{-2}$, $\boldsymbol{f}(\boldsymbol{v}) = \boldsymbol{v}^{-3}$, $\boldsymbol{f}(\boldsymbol{v}) = \log \boldsymbol{v}$, $\boldsymbol{f}(\boldsymbol{v}) = (\log \boldsymbol{v})^{-1}$.

Fitting procedure To determine the best choice for $\boldsymbol{f}$, we fit each possible function and evaluate the goodness of fit. As mentioned in Escudero Sanchez et al. [16], we use the same delineation for all samples from the same patient, implying we do not capture the VOI variability when factors are changed. Hence, we combine all observations for the same patient, summarizing the feature variability through the mean $\{\boldsymbol{\mu}_i\}$ and standard deviation $\{\boldsymbol{\sigma}_i\}$. We estimate the values of the corresponding β_1, β_0 through maximum likelihood estimation, assuming a Gaussian distribution around each observation. This comes out to:

$$\hat{\beta}_1, \hat{\beta}_0 = \arg\max_{\beta_1, \beta_0} \mathcal{L}(\{\boldsymbol{\mu}_i\}, \{\boldsymbol{\sigma}_i\} | \beta_1, \beta_0) = \arg\max_{\beta_1, \beta_0} \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\boldsymbol{\sigma}_i} e^{-\frac{(\boldsymbol{\mu}_i - \beta_1 f(\boldsymbol{v}_i) - \beta_0)^2}{2\boldsymbol{\sigma}_i^2}} \tag{3.2}$$

In practice, we fit this using a Weighted Least Squares (WLS) regression, which we show is equivalent to the above equation in App. A.3. We then compare the resulting fits and take the one with the lowest statistic value upon performing a χ^2 goodness-of-fit test:

$$\hat{\boldsymbol{f}} = \arg\min_{\boldsymbol{f}} \chi^2_{\boldsymbol{f}} = \arg\min_{\boldsymbol{f}} \sum_{i=1}^{N} \left(\frac{\boldsymbol{\mu}_i - \boldsymbol{f}(\boldsymbol{v}_i)}{\boldsymbol{\sigma}_i} \right)^2 \tag{3.3}$$

Removing the relationship Finally, to remove the discovered relationship, we apply the inverse of $\hat{\boldsymbol{g}}$: $\boldsymbol{x}_{\text{norm}} = \frac{\boldsymbol{x} - \hat{\beta}_0}{\boldsymbol{f}(\hat{\boldsymbol{v}})}$. We disregard the β_1 factor, as for our purposes the features are scale-invariant. This step ensures that the most influential effect we discover is isolated. It does not guarantee that we have removed the entire dependency of the feature on the VOI, due to lower-order factors, but is largely sufficient. Finally, we standardize the features to a mean of 0 and standard deviation of 1 to ensure numerical stability.

3.3 Model discrimination analysis

In order to prove the ability of certain radiomic features to detect differences between different PDX models, we measure how each improves a classifier that is trained on the given dataset. However, due to the low number of samples available and high cross-correlation of the entries, we first aim to reduce the feature space, which is originally of size 93. To this end, we develop a 2-step feature reduction pipeline, constituting of: 1) Removing features which show no discriminative power, and 2) Discarding highly correlated features. We then fit a classifier on the reduced feature set and determine the importance of each one in an appropriate manner. We perform each component separately in a cross-validation setting and observe whether different subsets result in the same features being discarded.

3.3.1 Statistical feature reduction

Removing non-discriminative features The first step of the feature reduction is to remove any features that have no discriminative power. We separate each feature into 2 groups corresponding to observations from the basal and luminal models, and aim to determine whether there is a statistically significant difference in their distributions. We first apply the Kruskal-Wallis test [37] to determine statistically insignificant features. We then apply the Benjamini-Hochberg correction [38] to account for the false discovery rate of multiple comparisons.

Kruskal-Wallis test The Kruskal-Wallis test is a statistical test that can differentiate samples from different distributions. It is a non-parametric method that can be considered as one-way ANOVA on the feature rank $\boldsymbol{r}$. Each feature rank is associated to the corresponding sample, which belongs to a "group" i with n_i observations. Since we want to detect the sensitivity with respect to the model, we group our observations in 2 classes - luminal and basal. The formula for the statistic can be found below.

Definition 3.3.1 The *Kruskal-Wallis statistic* for a set of observations $\boldsymbol{f}$ with an associated rank $\boldsymbol{r}$, such that any observation is associated with a group i is:

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}$$

, where N is the total number of observations g is the number of groups, n_i is the number of observations of group i, $\bar{r}_i$ is the mean rank of observations in group i, and $\bar{r}$ is the global mean.

P-values can be computed by looking up probability distribution tables, such as the one provided by Meyer and Seaman [39] can cover up to a total of 105 groups. The nature of this test allows us to extend our methodology to even more subtypes of breast cancer with similar success. In our work, we utilise the implementation provided by the *scipy* library [40]. We compute the p-values for all entries and all features, and then use the Benjamini-Hochberg correction up to remove irrelevant features up to a given false discovery rate threshold.

Benjamini-Hochberg correction The Benjamini-Hochberg procedure [38] is a method to control the false discovery rate when performing multiple hypothesis tests. Statistical chance implies we have to account for randomness in the obtained p-values. , removing any features above an adjusted p-value of $\alpha = 0.25$, which corresponds to an expected 25% false discovery rate. The technical details of the procedure are given below:

Definition 3.3.2 The *Benjamini-Hochberg correction* accounts for an expected false discovery rate of α by rejecting all p-values above a threshold τ. τ can be determined as:

1. Let the sorted p-values be $\boldsymbol{p}$ and their respective ranks be $\boldsymbol{r}$.
2. Compute adjusted p-values $\boldsymbol{p}'_i = \frac{N}{r_i}\boldsymbol{p}_i$
3. Set $\tau = \boldsymbol{p}'_k$, where $k = \arg\min_i \boldsymbol{p}'_i > \alpha$
4. Reject all hypotheses such that $\boldsymbol{p}'_i \geq \tau$.

Correlation-based filtering We now proceed by removing highly-correlated features, as keeping any pair of such features will essentially add a degree of freedom for the classifier to fine-tune its predictions, causing significant overfitting. To this end, we *Repeated Measures*

Correlation (RMCorr) statistic [41], as implemented by the *pingouin* package [42]. A crucial difference between RMCorr and the commonly used Pearson correlation is that the measures are adjusted for each patient. This circumvents issues described by Simpson's paradox [43] where categorical factors may hide underlying relationships between the features.

Definition 3.3.3 The *Repeated Measures Correlation* coefficient r_{RM} between features $\boldsymbol{x}^1, \boldsymbol{x}^2$ with $N_1, N_2, \ldots, N_P$ observations for P different "participants", such that $\boldsymbol{x}^k = \{\boldsymbol{x}^k_{ip}\}_{p\in[1,\ldots,P], i\in[1,\ldots,N_p]}$, is a measure of the best fit of the equation:

$$\hat{\boldsymbol{x}^1_{ip}} = \bar{\boldsymbol{x}^1_p} + \beta_p + \beta(\boldsymbol{x}^2_{ip} - \bar{\boldsymbol{x}^2_p}) \tag{3.4}$$

Where β and $\{\beta_p\}_{p\in[1,\ldots,P]}$ are free parameters. The coefficient itself is then computed as:

$$r_{\text{RM}}(\boldsymbol{x}^1, \boldsymbol{x}^2) = (-1)^{\mathbb{1}_{\beta>0}} \sqrt{\frac{SS_{\text{reg}}}{SS_{\text{reg}} + SS_{\text{err}}}} \tag{3.5}$$

With $SS_{\text{reg}} = \sum_{p=1}^{P} \sum_{i=1}^{N_p} (\hat{\boldsymbol{x}^1_{ip}})^2$ and $SS_{\text{err}} = \sum_{p=1}^{P} \sum_{i=1}^{N_p} (\hat{\boldsymbol{x}^1_{ip}} - \boldsymbol{x}^1_{ip})^2$

We iterate through all pairs of features $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ and compute the corresponding $r_{\text{RM}}(\boldsymbol{x}_i, \boldsymbol{x}_j)$. If the features are highly correlated ($r_{\text{RM}}(\boldsymbol{x}_i, \boldsymbol{x}_j) > \alpha$) for some predetermined α, we take the one with a higher discriminative power. We can achieve this by training three classifiers - Random Forest [44], Gradient Boosting Classifier [45] and Support Vector Machine [46], on a single feature and take the average accuracy across them. We then retain only the feature with the higher such score across each pair. Relevant hyperparameters for all classifiers can be found in Table A.1.

3.3.2 Forward selection

As part of our investigation into the methodology's robustness, we considered using *forward selection* in place of the statistical methods described above.

Definition 3.3.4 *Forward selection* [47] is a subset of stepwise regression and is used in classical machine learning for isolating statistically significant features. One usually starts with an empty model, iteratively adding the most effective feature through a selected criterion. Each feature is selected by exhaustion, being added to the previous model. The process terminates either when no further improvement can be achieved, or upon reaching a certain threshold.

In particular, we utilise the `scikit-learn` [48] implementation, with further technical details given in Ferri et al. [49]. We elect to keep the correlation-based filtering component, as we would like to isolate as many independent features as possible.

3.3.3 Feature importance

For determining the importance of each feature, we first fit a classifier on the reduced feature set. In particular we consider the tree-based Random Forest and Gradient Boosting Classifier, as well as the simpler Logistic Regression. We can easily determine the feature importance of the latter by examining the coefficients of the fit, with a higher absolute value implying a higher importance. While there is no trivial way to determine the value of a feature in tree-based classifiers, we utilise the SHAP values [50], which are approximation of the Shapley scores

that originate from a game-theoretical context [51, 52]. The SHAP values we derive assume a tree-based model, and are also performed in a cross-validated setting, taking into account the model's performance *only* on the test set. The model hyperparameters were determined through a grid-search using 5-fold cross-validation by optimizing the accuracy of the validation set.

Chapter 4

Results

In this chapter, we discuss the results and conclusions from the application of our methodology. We first present our findings on replicating the original work, detailing any significant differences and how they affect the conclusions of the paper. We then present ablation studies on each component, evaluating changes in the final or intermediate results that depend on our component choices.

4.1 Replication

We begin by discussing our replication of the original work. We followed the same methodology, while using an original implementation that focuses on replicability through using standard libraries and control of random number generation.

4.1.1 Sensitivity analysis

Main study We first perform the sensitivity analysis in the same setting as the prior work - removing one luminal model subject, and applying a balanced ANOVA. We plot a bar chart for the contribution of all factors for each feature, and compare it to the similar one from the original work [16] in Fig. 4.1. We observe the same 4 features with a notable sensitivity ($\eta^2 > 0.4$) to the underlying model - FO 10 Percentile, FO Kurtosis, FO Skewness, GLCM IDN, NGTDM Coarseness. Most other first-order features were highly sensitive to the reconstruction algorithm, or in the case of texture-based features - to the number of gray levels. Following Escudero Sanchez et al. [16], we then standardize over these two factors to eliminate the respective dependencies.

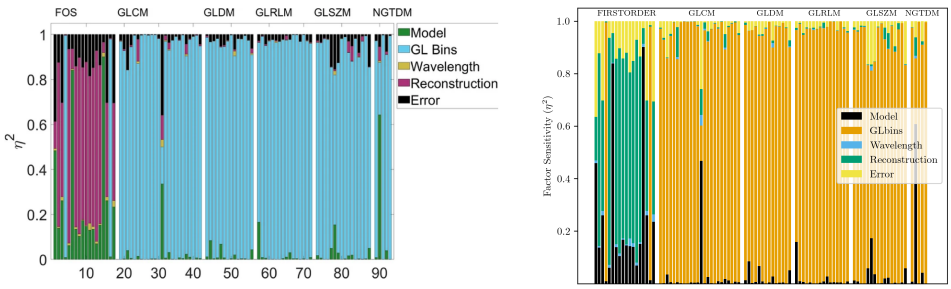


Figure 4.1: A comparison of the factor contribution between the original study (left) and our reproduction (right). The left figure was taken from Escudero Sanchez et al. [16].

Standardized gray levels When gray levels were standardized, as expected first-order features are not greatly affected, while the η^2 values for texture-based features vary plenty due to their high reliance on the gray level number. While there is no single optimal level observed, visibly we can conclude that anywhere between 32 and 128 bins allows for the most sensitive features, which also corresponds to the optimal value of 40 found in Escudero Sanchez et al. [36]. We include a side-to-side comparison for $N_g = 32, 64$ between our studies and the original work in Fig. 4.2, while the rest can be found in App. B.1.

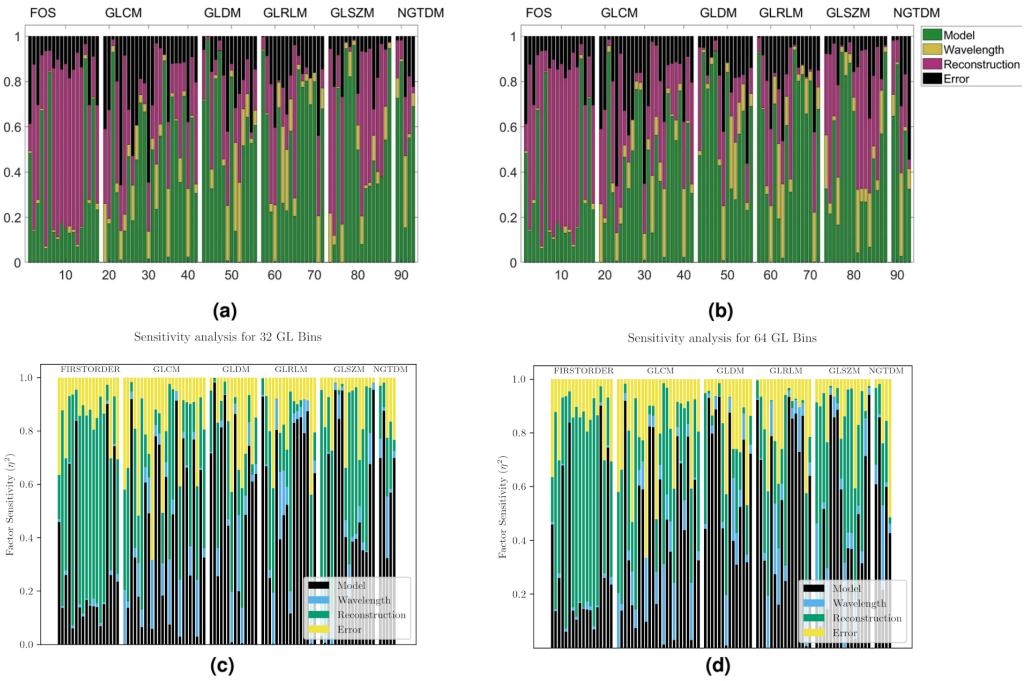


Figure 4.2: A comparison of the factor contribution when standardizing for GL bins (32 and 64 bins shown) between the original study - (a), (b) respectively and our reproduction - (c), (d) respectively. The top figure was taken from Escudero Sanchez et al. [16]

Standardized reconstruction method Similar to how first-order features were mostly unaffected when standardizing the gray levels, fixing the reconstruction method barely changes the η^2 values for texture-based features. On the other hand, we confirm Escudero Sanchez et al. [16]'s conclusions that regardless of the algorithm, most first-order features become highly sensitive to the tumour model, as seen in Fig. 4.3. The implications of this finding is that upon using the same reconstruction method, the first-order feature are likely to have a much higher predictive utility.

K-fold cross validation Finally, we perform a 5-fold cross validation, removing 3 luminal and 2 basal model subjects from each iteration and evaluating the appropriate mean and standard deviation of the measured η^2. We observe values agreeing with the original work in Table 4.2, implying that the FO Skewness feature is more generalisable under different samples. In the supplementary pipeline, we support a script that derives in the case of needing further inspection. Furthermore, we

Table 4.1: Observed mean and standard deviation for other relevant features.

Feature	**RP** $\mu \pm \sigma$	**CoV**
FO 10 Percentile	0.47 ± 0.03	0.07
GLCM IDN	0.43 ± 0.32	0.74
NGTDM Coarseness	0.56 ± 0.07	0.13

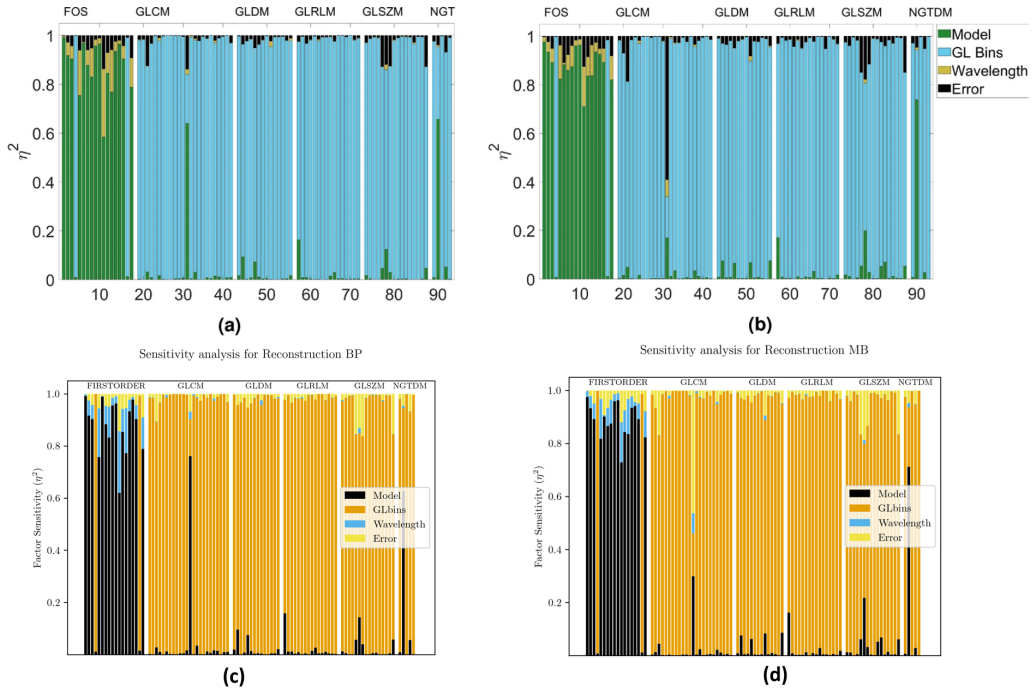


Figure 4.3: A comparison of the factor contribution when standardizing for reconstruction method (Backprojection vs Model-based) between the original study - (a), (b) respectively and our reproduction - (c), (d) respectively. The top figures were taken from Escudero Sanchez et al. [16]

Table 4.2: Comparison between the observed mean and standard deviation for the measured η^2 on the FO Kurtosis and FO Skewness features between the original (OG) work and our reproduction (RP).

	Kurtosis				**Skewness**			
Factor	**RP** $\mu \pm \sigma$	**RP CoV**	**OG** $\mu \pm \sigma$	**OG CoV**	**RP** $\mu \pm \sigma$	**RP CoV**	**OG** $\mu \pm \sigma$	**OG CoV**
Model	0.78 ± 0.19	0.24	0.73 ± 0.16	0.22	0.89 ± 0.07	0.08	0.86 ± 0.07	0.08
GL Bins	0.00 ± 0.00	N/A	0.00 ± 0.00	N/A	0.00 ± 0.00	N/A	0.00 ± 0.00	N/A
Wavelength	0.01 ± 0.01	1.13	0.02 ± 0.02	1.12	0.01 ± 0.01	1.12	0.03 ± 0.02	0.74
Reconstruction	0.14 ± 0.12	0.82	0.15 ± 0.08	0.54	0.06 ± 0.04	0.61	0.07 ± 0.03	0.42
Error	0.07 ± 0.06	0.88	0.10 ± 0.07	0.70	0.03 ± 0.02	0.50	0.10 ± 0.03	0.57

include the statistics for the model sensitivity on the other 3 features reported as sensitive above in Table 4.1, observing that the GLCM IDN feature exhibits unusual variability. These statistics indicate the feature exhibits some signal (high mean), but is also highly unreliable (high variance). It is further notable that this particular feature is not among the top 9 in any of the model discrimination analysis tests performed for random forest classifiers, despite being in the reduced feature set, further showing its unreliability.

4.1.2 Model discrimination analysis

In this section, we replicate the results of the model discrimination analysis in Escudero Sanchez et al. [16]. The original work provides the VOI-normalized features used in their paper, which deviates from what we derive, as explored in Sec. 4.2.2. Here we aim to identify the usability of the methodology regardless of the normalization procedure. To this end, we make use of the preprocessed data, provided in the respective GitHub repository.

Statistical feature reduction We first begin with removing the features which are likely to not have sufficient signal for differentiating between models. We apply the Kruskal-Wallis test on the two groups of features and apply the Benjamini-Hochberg correction with $\alpha = 25\%$. We obtain an **exact** replication of the original work, with the obtained p-values listed in App. B.3. This is possible due to no random effects. We further perform an ablation study on the choice of the Kruskal-Wallis test in Sec. 4.2.4.

Removing highly-correlated features We now remove highly correlated feature using the RMCorr [41] measure, as described in Sec. 3.3.1. Before proceeding, we highlight the importance of standardizing the data before this step. Not doing so causes numerical instabilities within the framework due to the varying orders of magnitude of the normalized features, which can vary between $\approx 10^{-15}$ to $\approx 10^{20}$. Comparing to the supplementary material of the original work, we observe some differences in the correlation matrix (App. B.2). It is possible that the cause is the underlying numerical instabilities, however, the effects of the deviations are not significant. We discover that after performing the feature selection pipeline, we come out with 26 different features, instead of the proposed 27, which we list below:

- FO 10Percentile, Entropy, Kurtosis, Minimum, RootMeanSquared (or 90Percentile, both of which are equivalent in terms of correlation and discriminative properties), Skewness
- GLCM ClusterProminence, ClusterShade, Idmn, Idn, Imc2, MCC, SumSquares
- GLDM DependenceNonUniformityNormalized, LargeDependenceEmphasis, SmallDependenceHighGrayLevelEmphasis
- GLRLM GrayLevelNonUniformity, GrayLevelNonUniformityNormalized, GrayLevelVariance, LowGrayLevelRunEmphasis
- GLSZM GrayLevelNonUniformity, LargeAreaEmphasis, LowGrayLevelZoneEmphasis
- NGTDM Busyness, Coarseness, Strength

There is a significant overlap, with the only difference being the GLSZM SizeZoneNonUniformityNormalized feature, which was deemed highly correlated with the GLCM JointEntropy. We further perform an ablation study on the selection of the correlation threshold α_{corr} in Sec. 4.2.6.

Feature Importance We conclude the reproduction part by observing how the changes in the feature reduction have affected the inferred feature importance from the SHAP plots. We present the top 9 features and their respective scores and score distributions in Fig. 4.4. We achieve agreement on 8/9 features, retrieving GLCM Imc2 instead of GLRLM LowGrayLevelRunEmphasis. More notably, the distributions seen on the beeswarm plots implies similar SHAP values between corresponding features, implying they affect the prediction similarly. This implies that these features are more robust to small changes in the feature set, which will be further explored in Sec. 4.2.5 and Sec. 4.2.7.

4.2 Ablation studies

Because we wish to isolate the effects of each component in the pipeline, and whether the final result is affected by any changes, we perform ablation studies on the majority of the steps. We primarily focus on the model discrimination analysis, as it contains more components, which may not be robust to changes in hyperparameters or methodology. We also perform a small modification of the ANOVA performed in the sensitivity analysis, allowing for unbalanced

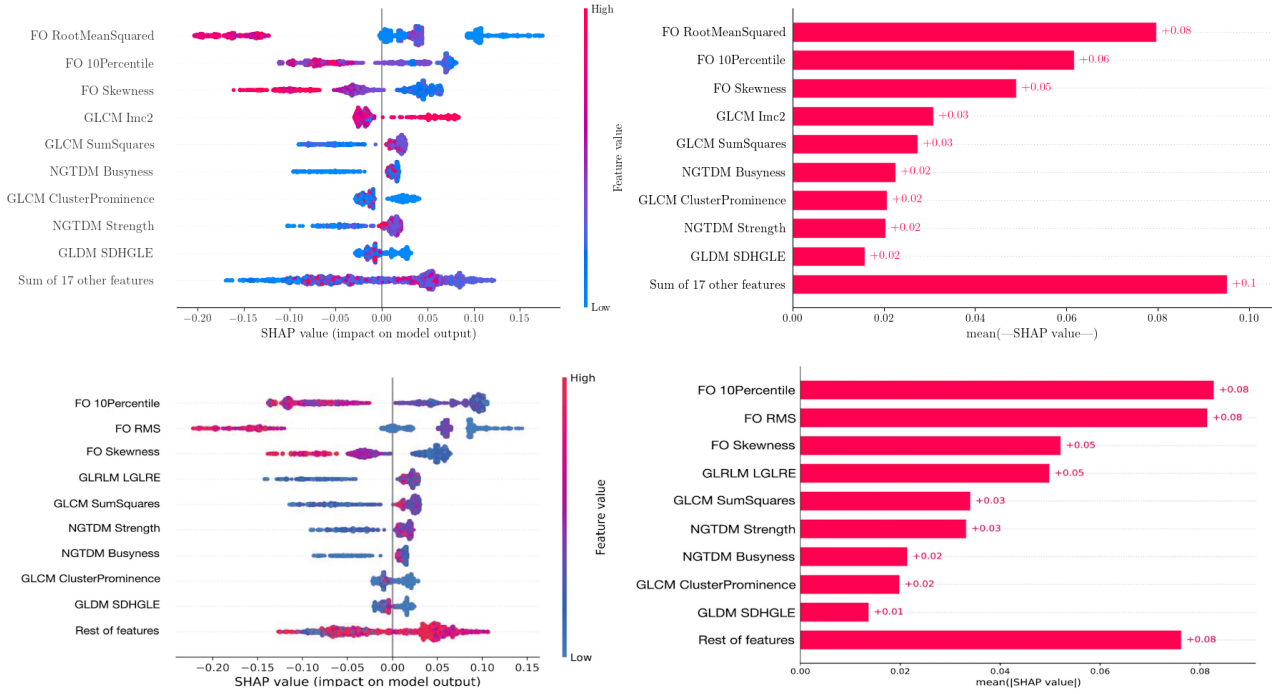


Figure 4.4: A comparison of the SHAP plots between the original study (bottom) and our reproduction (top).

factors, in order to observe if there are any significant differences to including the extra luminal specimen.

4.2.1 Effect of balanced ANOVA

Escudero Sanchez et al. [16] performed the sensitivity analysis by removing one luminal model subject in order to ensure both classes are equally represented. However, the generalized formula we present in Sec. 3.1 shows that this is not necessary to perform ANOVA. We present the resulting effects in Fig. 4.5. We notice the only significant change occurs in the feature GLCM Inverse Difference Normalized (IDN), which exhibits a large deviation considering the one-sample difference, as previously noted in Sec. 4.1.1. We can therefore conclude that the sensitivity analysis is robust with respect to *slight class imbalance*. The implications of this fact are covered more thoroughly in Chapter 5.

4.2.2 Effect of VOI normalization prior knowledge

We now proceed to discussing the effects of different components of the model discrimination analysis, starting with the VOI normalization procedure. Due to the limited nature of the dataset, Escudero Sanchez et al. [16] utilise the same functional transformations that were found in the original procedure [36]. However, since we do not have access, we train and apply the training procedure on the limited data from this study. While we do observe significant differences in the process of feature reduction, we focus on the final results, for which we include the SHAP plots in Fig. 4.6. While we have an overlap in 6/9 of the top 9 features, there are significant differences in both distributions and relative rank, further showing the inconsistencies. Through these results we emphasize the need for a standardized normalization procedure, especially in cases where radiomic features are considered for clinical use.

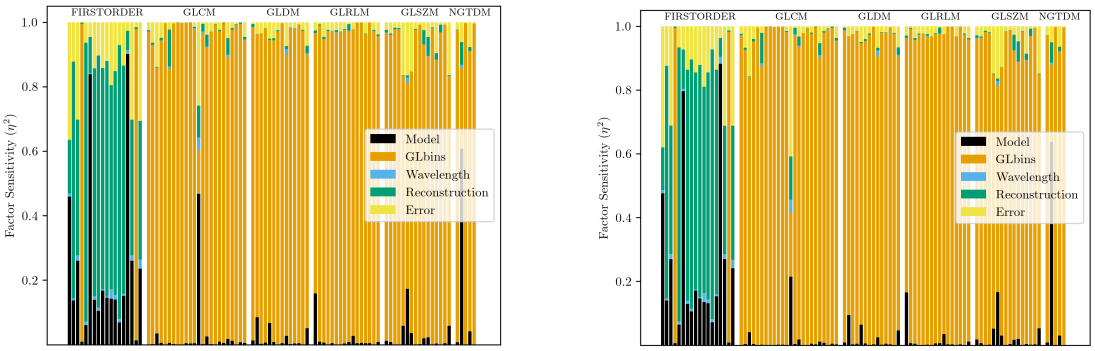


Figure 4.5: A comparison of the factor contribution of balanced (left) and unbalanced (right) ANOVA.

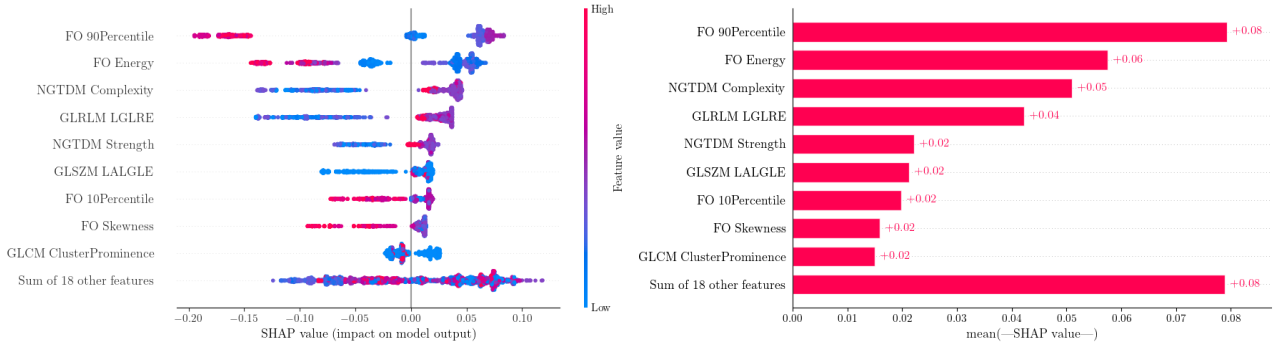


Figure 4.6: A beeswarm and bar-chart for the top 9 features, as derived by the SHAP values on a Random Forest classifier on the features normalized in this work.

4.2.3 Effect of feature reduction method

We hence proceed with using the provided features, which were preprocessed as part of the original work. This section focuses on the choice of feature reduction methodology using statistical methods, instead approaching the task using forward selection, as described in Sec. 3.3.2. We obtain 46 features after the forward selection step, with only 23 remaining after removing highly-correlated features. We can observe the top 9's importance in Fig. 4.7. We do in fact have a large overlap in the features and their behaviour. However, considering no features from the NGTDM or GLRLM feature groups passed through the forward selection, it is likely that the method is highly unstable for a limited dataset. The results obtained here are more likely showing the robustness of the remaining components of the pipeline, but it is reassuring that a radical component change does not affect the results significantly.

4.2.4 Effect of statistical test choice

We apply the same pipeline as done in the main study, substituting for the use of the Kolmogorov-Smirnov test [53]. We notice that the test is far less powerful, unable to detect differences in distribution for most features, **only removing 6** features through the Benjamini-Hochberg correction, as seen in Table B.4. Hence, most of the feature selection is achieved through the removal of highly correlated features, which will be explored further in the next section. We can conclude that the Kruskal-Wallis test is more appropriate in this situation due its higher discriminative power when a lower number of samples is involved.

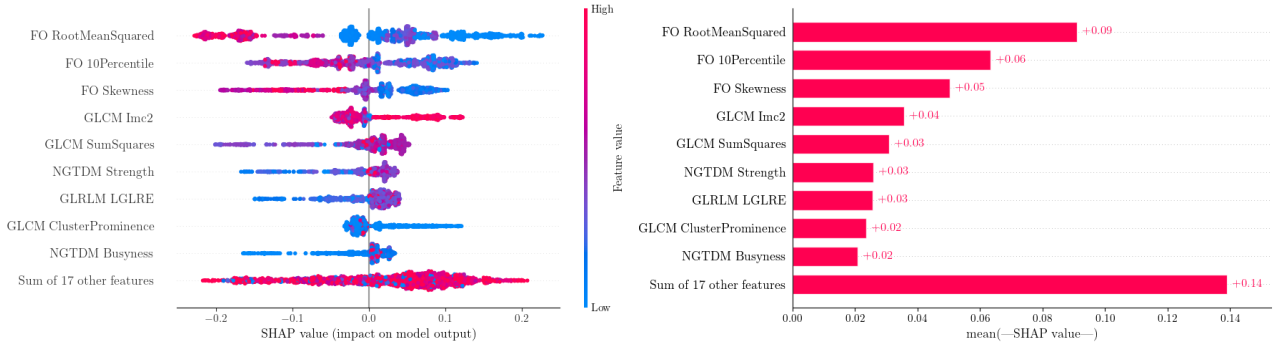


Figure 4.7: A beeswarm and bar-chart for the top 9 features, as derived by the SHAP values on a Random Forest classifier after performing forward selection.

4.2.5 Effect of cross-validation

We now perform the feature importance step of the pipeline in a cross-validated setting using the inferred optimal hyperparameters. We observe a quite similar ranking of the features compared to what we obtained in Sec. 4.1.2, as shown in Fig. 4.8. Out of the previously observed top 9 features, we discover that only GLDM SmallDependenceHighGrayLevelEmphasis is replaced by GLRLM LowGrayLevelRunEmphasis (which are not significantly correlated with a correlation coefficient of ≈ -0.13). Considering no major deviations were found, and the fact that cross-validation is usually more generalisable than evaluating on the training set, we will proceed with using cross-validation for the following ablation studies.

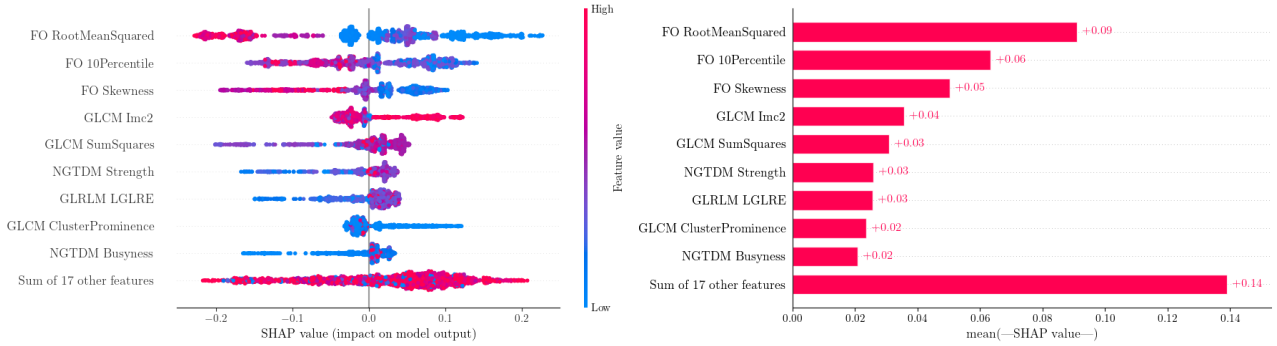


Figure 4.8: A beeswarm and bar-chart for the top 9 features, as derived by the SHAP values on a Random Forest classifier using cross-validation.

4.2.6 Effect of correlation threshold

We perform an investigation of how the correlation threshold α_{corr} affects our final conclusions. We tried an additional 3 values for $\alpha_{corr} = 0.8, 0.7, 0.6$ to test a large range of reasonable values. For each of them, we retain 19, 9 and 8 features respectively after reduction is performed. Thus we can conclude that $\alpha_{corr} = 0.7, 0.6$ are overly pessimistic values, as the predictive capabilities of the logistic regression drop to a test accuracy of 63.2% and 59.7% respectively from 69.9% in a cross-validated setting. We used the logistic regression as our benchmark, as it is the model that is least prone to over- or underfitting. We present the resulting top 9 features in Fig. 4.9. The main takeaway is that the majority of the features remain the same, with the most prominent change being FO Skewness being replaced with FO Kurtosis, which were deemed to

be correlated with $\rho_{corr} > 0.8$. We do, however, observe an increase in reliance in the best 2 features, which is a result of the more conservative feature reduction where the model cannot use the orthogonal information from the discarded features.

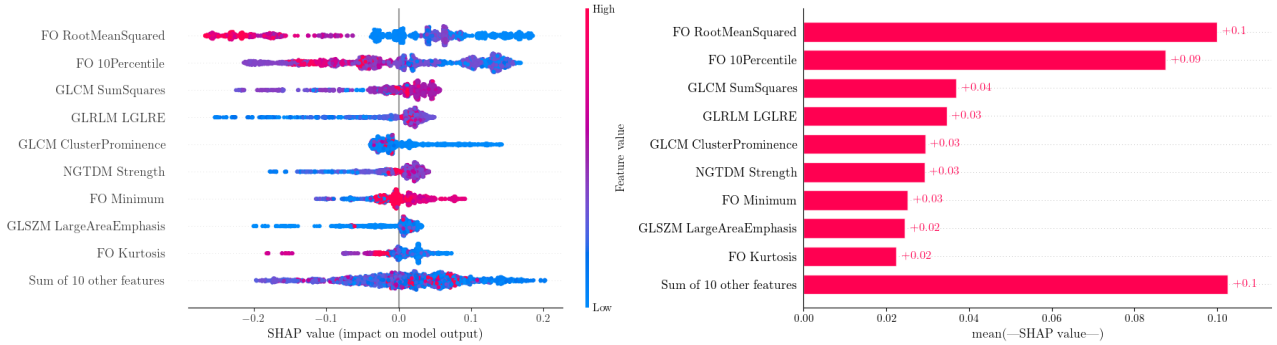


Figure 4.9: A beeswarm and bar-chart for the top 9 features, as derived by the SHAP values on a Random Forest classifier after applying feature reduction with $\alpha_{corr} = 0.8$.

4.2.7 Effect of model choice

Finally, we inspect the effect of our model choice. Up until now, we primarily determined feature importance through the SHAP values of a random forest classifier. In this section, we further evaluate the features through a gradient boosting classifier and the logistic regression, using the hyperparameters from the grid search (Table A.1). We produce the SHAP plot for the gradient boosting classifier in Fig. 4.10. Given that it barely relies on more than 2 features, as well as that we achieve a perfect 100% accuracy in both the training and test set, it is likely that this model is underdetermined, and unfortunately cannot be used for this study. We then consider

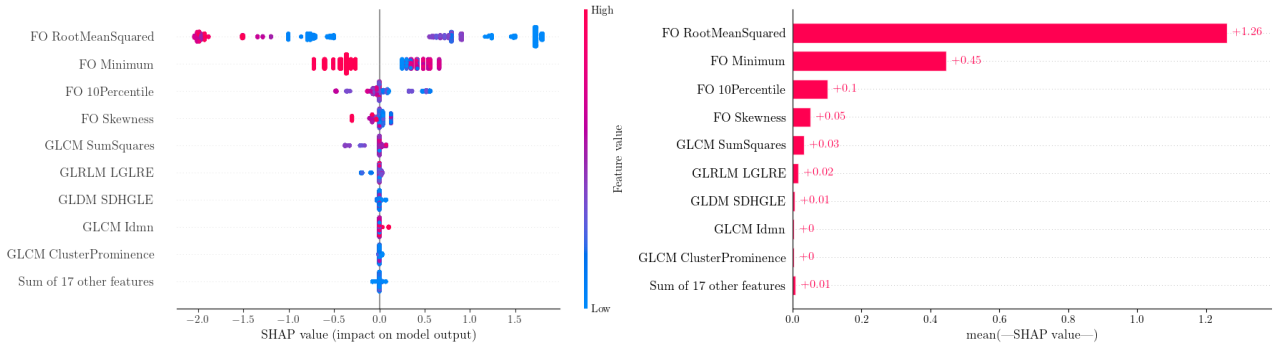


Figure 4.10: A beeswarm and bar-chart for the top 9 features, as derived by the SHAP values on a gradient boosting classifier

the logistic regression coefficient as a measurement of feature importance, visualized in Fig. 4.11. We notice the results are more in line with what we expect from the sensitivity analysis, likely due to the embedded linear relationships. Namely, we observe that FO Skewness, FO Kurtosis, GLCM IDN, FO 10Percentile are all among the top 10. We do not observe a complete overlap with the random forest method, as we do not capture the GLCM Imc2, GLRLM LGLRE and NGTDM Strength features, with most of the rest having a significant deviation in their relative importance rank. Therefore, we can likely conclude that the nature of the predictive model is a

significant factor, with some features showing natural predictive power regardless of this choice.

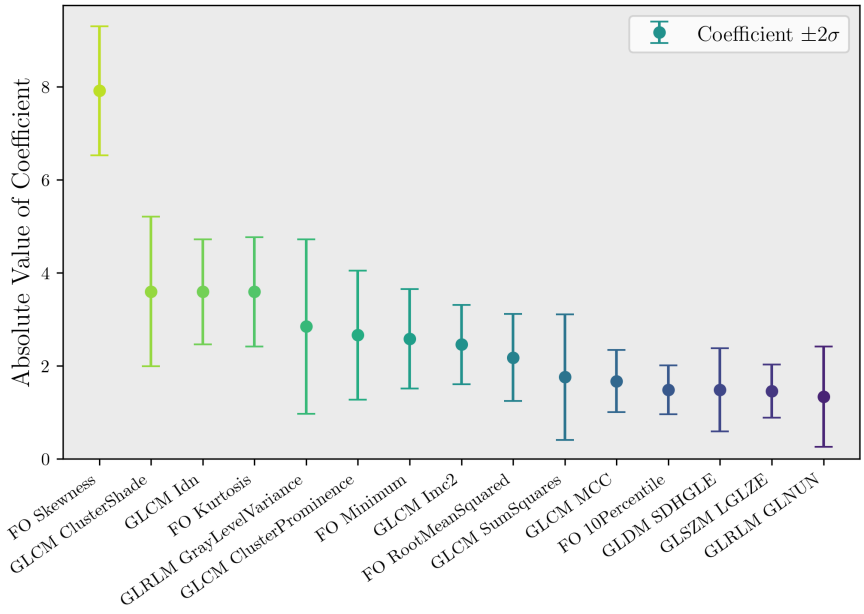


Figure 4.11: The absolute values of the coefficients of the logistic regression, with errors determined by the covariance matrix.

> **Remark:** For the sake of determining importance through the logistic regression coefficients, all features must have **the same variance**, so that the coefficients are comparable. We achieve this through standardization to a $\sigma^2 = 1$.

Chapter 5

Conclusions, Discussion and Further work

In this section, we discuss the implications of our replication and the tangential extensions to the original work. As a preface, we were successfully able to replicate with high degree of overlap, the original findings of Escudero Sanchez et al. [16]. We further support or pose further questions to the model discrimination analysis using ablation studies on each component. We aim to show what directions for optimization might be feasible, in order to make radiomic features more viable for clinical use.

Sensitivity analysis First of all, we performed the sensitivity analysis (Sec. 4.1.1) with near-identical results to the original work. The point of this step is to create a descriptive set of measurements for investigating certain radiomic features and how they might be affected under different settings, which have not been sufficiently investigated in related work. Through this analysis, we reaffirm the discovery that several first-order feature are particularly sensitive to the underlying cancer model, namely the *Kurtosis*, *Skewness* and the 10^{th} *Percentile*. The latter two are also shown to be stable under distribution variability through k-fold cross-validation.

We also demonstrated the grey levels and as reconstruction algorithm as significant factors for different groups of radiomic feature, implying standardization may be required for the features to be more discriminative. Furthermore, we showed that removing a single sample to achieve a balanced distribution may not be necessary, as similar results can be obtained through a slight modification that allows for the use of the entire dataset. This statement holds for slightly unbalanced data, as this will imbue a prior distribution into the sensitivity measurement, which can be useful under particular assumptions. However, in the case of an abundant or rare class, this methodology may become inappropriate.

Model discrimination analysis With regards to the model discrimination analysis, we reaffirm the conclusions made by Escudero Sanchez et al. [16], achieving exact reproduction with the omission of a single feature at the final step. This, however, does not change the conclusion in a significant way, except that we discover the GLCM Imc2 feature among the most powerful ones. This claim is further substantiated in the ablation studies, where the feature remains among the most discriminative.

Our main contribution was to evaluate the framework under different settings, starting with the data pre-processing. We understand that the particular method of normalization, which takes into account the VOI, must be performed under a standardized setting, as this step is crucial to correctly isolating the predictive power of each feature. This can perhaps be achieved

under a Bayesian framework to allow for the accumulation of knowledge, as the datasets grow larger.

We further changed each component of the pipeline, while keeping the rest constant. First of all, we discovered that the choice of statistical test is appropriate, as the Kruskal-Wallis test was more powerful, compared to the more commonly used Kolmogorov-Smirnov test, in addition to also being extendable to multiple classes. Classical machine learning methods, such as the forward selection also performed similarly, but was deemed unstable due to being biased towards certain groups of features. We further showed the choice for a correlation threshold α_{corr} is most likely optimal enough so that a sufficient number of features are included in the final evaluation. We reaffirm the need to perform the analysis under a cross-validated setting, as that ensures better generalisability. Even given concerns for the small nature of the dataset, using a 5-fold validation procedure we were able to observe similar results with lower uncertainty. This study used a random forest framework to determine the conclusions presented above, so we aim to verify our knowledge with respect to this factor. Although a more powerful model like the gradient boosting classifier tends to overfit with a small dataset and is not suitable for measuring importance, we opted for the simpler logistic regression model, which is also not tree-based. We discovered that while the behaviour is quite different, it is heavily linked to the sensitivity analysis in terms of distinguishing features, but also has significant overlaps with aforementioned results. It is highly likely that this is caused by the vastly different model structures, and the fact that a feature cannot be a lone predictor, and must be evaluated in combination with others.

Finally, we conclude that under almost any setting, the following features are found to have a high predictive potential:

- FO Skewness
- FO 10Percentile
- FO 90Percentile/RMS
- GLCM SumSquares
- NGTDM Strength
- GLCM ClusterProminence

Limitations and further work We were primarily limited by the small size of our dataset, featuring only 21 different subjects. A study on a much larger scale is necessary to provide further insight into topics discussed in this and the original works. This limitation imposes further biases in how we performed the analysis with sub-optimal data scientific techniques, through reusing the test and training set for further validation. Furthermore, we only apply the pipeline for differentiating between 2 of the 4 main subtypes of breast cancer. More diversity and potentially including negative information through a control group, will ensure the significance of our findings. Therefore, any conclusions made should be taken into account as avenues for further work. Finally, we do not take into account different delineations that can be performed when segmenting the tumour, meaning that a possible improvement into the pipeline can be to do so in order to further incorporate the shape features.

Conlusion To summarize, we were succesfully able to replicate the prior work [16], with very few conclusions being changed after performing the ablation studies. We have identified suitable radiomic features with discriminative power, as well as methods to evaluate them

through different classifier models. Finally, we emphasise the need for robustness within such a framework, ensuring that there are no unstable components within the pipeline.

Bibliography

[1] Tong Tong, Wenhui Huang, Kun Wang, Zicong He, Lin Yin, Xin Yang, Shuixing Zhang, and Jie Tian. Domain transform network for photoacoustic tomography from limited-view and sparsely sampled data. *Photoacoustics*, 19:100190, 2020.

[2] Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 74(3):229–263, 2024.

[3] Erasmo Orrantia-Borunda, Patricia Anchondo-Nuñez, Lucero Evelia Acuña-Aguilar, Francisco Octavio Gómez-Valles, and Claudia Adriana Ramírez-Valdespino. Subtypes of breast cancer. *Breast Cancer [Internet]*, 2022.

[4] Ozlem Yersal and Sabri Barutca. Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World journal of clinical oncology*, 5(3):412, 2014.

[5] Minghua Xu and Lihong V Wang. Photoacoustic imaging in biomedicine. *Review of scientific instruments*, 77(4), 2006.

[6] Paul Beard. Biomedical photoacoustic imaging. *Interface focus*, 1(4):602–631, 2011.

[7] Keerthi S Valluru and Juergen K Willmann. Clinical photoacoustic imaging of cancer. *Ultrasonography*, 35(4):267, 2016.

[8] Mohammad Mehrmohammadi, Soon Joon Yoon, Douglas Yeager, and Stanislav Y Emelianov. Photoacoustic imaging for cancer detection and staging. *Current Molecular Imaging (Discontinued)*, 2(1):89–105, 2013.

[9] Srivalleesha Mallidi, Geoffrey P Luke, and Stanislav Emelianov. Photoacoustic imaging in cancer detection, diagnosis, and treatment guidance. *Trends in biotechnology*, 29(5): 213–221, 2011.

[10] Michal R Tomaszewski and Robert J Gillies. The biological meaning of radiomic features. *Radiology*, 298(3):505–516, 2021.

[11] Janita E Van Timmeren, Davide Cester, Stephanie Tanadini-Lang, Hatem Alkadhi, and Bettina Baessler. Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights into imaging*, 11(1):91, 2020.

[12] Marius E Mayerhoefer, Andrzej Materka, Georg Langs, Ida Häggström, Piotr Szczypiński, Peter Gibbs, and Gary Cook. Introduction to radiomics. *Journal of Nuclear Medicine*, 61 (4):488–495, 2020.

[13] Alanna Vial, David Stirling, Matthew Field, Montserrat Ros, Christian Ritz, Martin Carolan, Lois Holloway, and Alexis A Miller. The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: a review. *Translational Cancer Research*, 7(3), 2018.

[14] Luca Brunese, Francesco Mercaldo, Alfonso Reginelli, and Antonella Santone. An ensemble learning approach for brain cancer detection exploiting radiomic features. *Computer methods and programs in biomedicine*, 185:105134, 2020.

[15] Shoshana B Ginsburg, Ahmad Algohary, Shivani Pahwa, Vikas Gulani, Lee Ponsky, Hannu J Aronen, Peter J Boström, Maret Böhm, Anne-Maree Haynes, Phillip Brenner, et al. Radiomic features for prostate cancer detection on mri differ between the transition and peripheral zones: preliminary findings from a multi-institutional study. *Journal of Magnetic Resonance Imaging*, 46(1):184–193, 2017.

[16] Lorena Escudero Sanchez, Emma Brown, Leonardo Rundo, Stephan Ursprung, Evis Sala, Sarah E Bohndiek, and Ignacio Xavier Partarrieu. Feasibility and sensitivity study of radiomic features in photoacoustic imaging of patient-derived xenografts. *Scientific Reports*, 12(1):15142, 2022.

[17] Idan Steinberg, David M Huland, Ophir Vermesh, Hadas E Frostig, Willemieke S Tummers, and Sanjiv S Gambhir. Photoacoustic clinical imaging. *Photoacoustics*, 14:77–98, 2019.

[18] Allan Rosencwaig and Allen Gersho. Theory of the photoacoustic effect with solids. *Journal of Applied Physics*, 47(1):64–69, 1976.

[19] Alexander Dima, Neal C Burton, and Vasilis Ntziachristos. Multispectral optoacoustic tomography at 64, 128, and 256 channels. *Journal of biomedical optics*, 19(3):036021–036021, 2014.

[20] Andrés Larroza, Vicente Bodí, David Moratal, et al. Texture analysis in magnetic resonance imaging: review and considerations for future applications. *Assessment of cellular and organ function and dysfunction using direct and derived MRI methodologies*, pages 75–106, 2016.

[21] Burak Koçak, Emine Şebnem Durmaz, Ece Ateş, and Özgür Kılıçkesmez. Radiomics with artificial intelligence: a practical guide for beginners. *Diagnostic and interventional radiology*, 25(6):485, 2019.

[22] Kaustav Bera, Nathaniel Braman, Amit Gupta, Vamsidhar Velcheti, and Anant Madabhushi. Predicting cancer outcomes with radiomics and artificial intelligence in radiology. *Nature reviews Clinical oncology*, 19(2):132–146, 2022.

[23] Hugues Benoit-Cattin. *Texture analysis for magnetic resonance imaging*. Texture Analysis Magn Resona, 2006.

[24] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973.

[25] MM Galloway. Texture classification using gray level run length. *Comput. Graph. Image Process*, 4(2):172–179, 1975.

[26] Guillaume Thibault, Jesus Angulo, and Fernand Meyer. Advanced statistical matrices for texture characterization: application to cell classification. *IEEE Transactions on Biomedical Engineering*, 61(3):630–637, 2013.

[27] Moses Amadasun and Robert King. Textural features corresponding to textural properties. *IEEE Transactions on systems, man, and Cybernetics*, 19(5):1264–1274, 1989.

[28] Chengjun Sun and William G Wee. Neighboring gray level dependence matrix for texture classification. *Computer Vision, Graphics, and Image Processing*, 23(3):341–352, 1983.

[29] Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.

[30] Alejandra Bruna, Oscar M Rueda, Wendy Greenwood, Ankita Sati Batra, Maurizio Callari, Rajbir Nath Batra, Katherine Pogrebniak, Jose Sandoval, John W Cassidy, Ana Tufegdzic-Vidakovic, et al. A biobank of breast cancer explants with preserved intra-tumor heterogeneity to screen anticancer compounds. *Cell*, 167(1):260–274, 2016.

[31] Stefan Morscher, Wouter HP Driessen, Jing Claussen, and Neal C Burton. Semi-quantitative multispectral optoacoustic tomography (msot) for volumetric pk imaging of gastric emptying. *Photoacoustics*, 2(3):103–110, 2014.

[32] M Li, Y Tang, and J Yao. Photoacoustic tomography of blood oxygenation: A mini review. photoacoustics 10 (2018): 65–73, 2018.

[33] Michaela Taylor-Williams, Graham Spicer, Gemma Bale, and Sarah E Bohndiek. Non-invasive hemoglobin sensing and imaging: optical tools for disease diagnosis. *Journal of Biomedical Optics*, 27(8):080901, 2022.

[34] R Schofield, L King, U Tayal, I Castellano, J Stirrup, F Pontana, James Earls, and E Nicol. Image reconstruction: Part 1–understanding filtered back projection, noise and image acquisition. *Journal of cardiovascular computed tomography*, 14(3):219–225, 2020.

[35] Lu Ding, Daniel Razansky, and Xose Luis Dean-Ben. Model-based reconstruction of large three-dimensional optoacoustic datasets. *IEEE transactions on medical imaging*, 39(9): 2931–2940, 2020.

[36] Lorena Escudero Sanchez, Leonardo Rundo, Andrew B Gill, Matthew Hoare, Eva Mendes Serrao, and Evis Sala. Robustness of radiomic features in ct images with different slice thickness, comparing liver tumour and muscle. *Scientific reports*, 11(1):8262, 2021.

[37] Patrick E McKight and Julius Najab. Kruskal-wallis test. *The corsini encyclopedia of psychology*, pages 1–1, 2010.

[38] JA Ferreira and AH Zwinderman. On the benjamini–hochberg method. 2006.

[39] J Patrick Meyer and Michael A Seaman. A comparison of the exact kruskal-wallis distribution to asymptotic approximations for all sample sizes up to 105. *The Journal of Experimental Education*, 81(2):139–156, 2013.

[40] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17 (3):261–272, 2020.

[41] Jonathan Z Bakdash and Laura R Marusich. Repeated measures correlation. *Frontiers in psychology*, 8:252904, 2017.

[42] Raphael Vallat. Pingouin: statistics in python. *J. Open Source Softw.*, 3(31):1026, 2018.

[43] Jan Sprenger and Naftali Weinberger. Simpson's paradox. 2021.

[44] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[45] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[46] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14:199–222, 2004.

[47] Michael Alin Efroymson. Multiple regression analysis. *Mathematical methods for digital computers*, pages 191–203, 1960.

[48] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12: 2825–2830, 2011.

[49] Francesc J Ferri, Pavel Pudil, Mohamad Hatef, and Josef Kittler. Comparative study of techniques for large-scale feature selection. In *Machine intelligence and pattern recognition*, volume 16, pages 403–413. Elsevier, 1994.

[50] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[51] Eyal Winter. The shapley value. *Handbook of game theory with economic applications*, 3: 2025–2054, 2002.

[52] Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *International conference on machine learning*, pages 9269–9278. PMLR, 2020.

[53] Vance W Berger and YanYan Zhou. Kolmogorov–smirnov test: Overview. *Wiley statsref: Statistics reference online*, 2014.

Appendix A

Deferred technical details

A.1 inVision 256-TF sample image

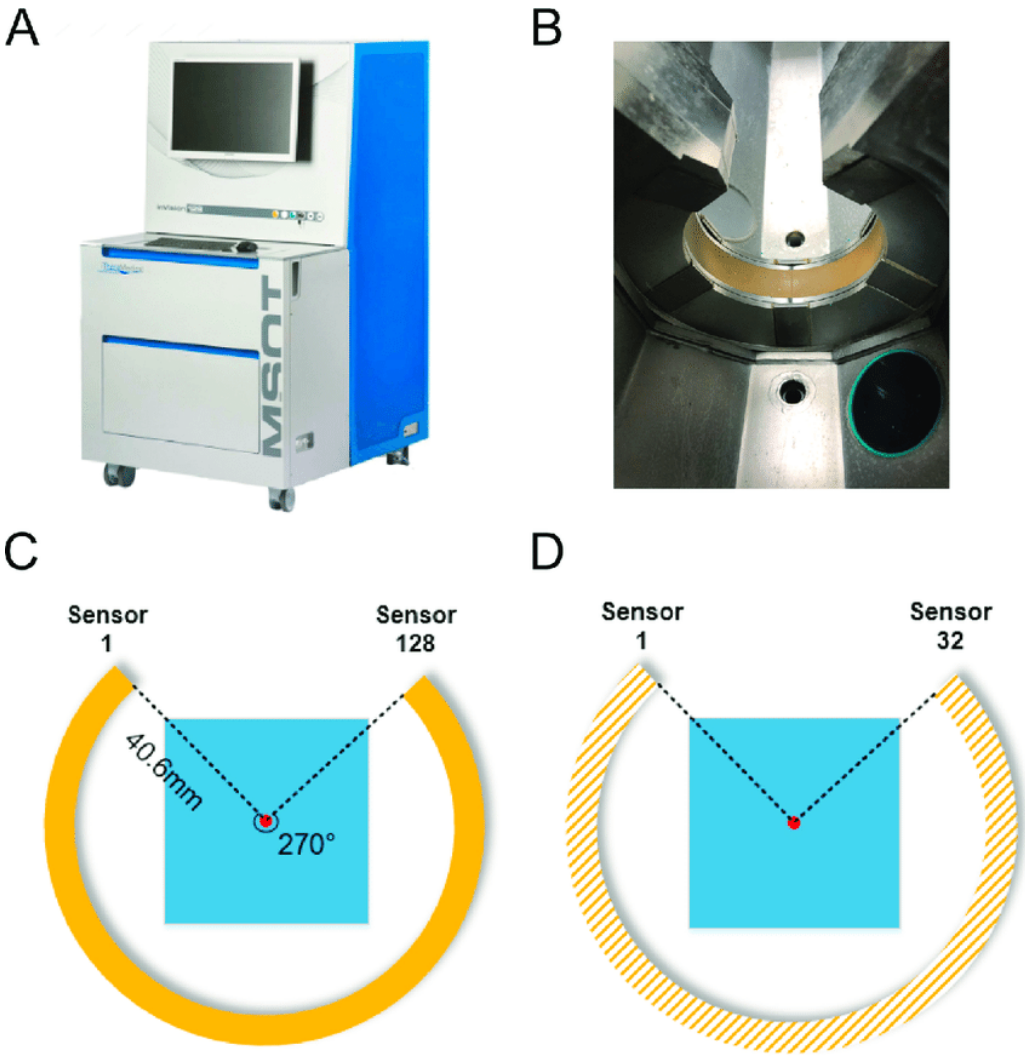


Figure A.1: Sample image of the inVision 128 scanner, which is visually identical to the inVision 256, albeit with a smaller number of sensors. (A) shows the device as seen from outside, while (B) shows the inside, specifically one of the transducers. (C) and (D) , on the other hand, show the schematics of the transducer. The image was provided by Tong et al. [1].

A.2 Ethical approval and licensing

This work, being based on the research done by Escudero Sanchez et al. [16], is under the same ethical and licensing guidelines. As described in the prior work, animal procedures were conducted in accordance with project and personal licenses, issued under the UK Animals (Scientific Procedures) Act from 1986. It was locally approved by the CRUK Cambridge Institute Animal Welfare and Ethical Review Board under compliance forms CFSB1567 and CFSB1979. Lastly, all animal methods and results are reported in accordance with the ARRIVE guidelines

A.3 Deferred proofs

Lemma A.3.1 The maximum likelihood estimation for the parameters β_1, β_0 for the below likelihood is equivalent to fitting performing a Weighted Least Squares (WSL) regression.

$$\mathcal{L}(\{\boldsymbol{\mu}_i\}, \{\boldsymbol{\sigma}_i\}|\beta_1, \beta_0) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\boldsymbol{\sigma}_i}} e^{-\frac{(\boldsymbol{\mu}_i - \beta_1 f(\boldsymbol{v}_i) - \beta_0)^2}{2\boldsymbol{\sigma}_i^2}} \tag{A.1}$$

Proof A.3.2 We begin by stating the optimization goal of the WLS regression. The WLS takes a set of observations $\boldsymbol{y}$, $\boldsymbol{x}$ and a set of weights $\boldsymbol{w}$ and seeks to minimize the following metric:

$$f(\alpha, \beta) = \sum_{i=1}^{N} \boldsymbol{w}_i (\boldsymbol{y}_i - \alpha \boldsymbol{x}_i - \beta)^2 \tag{A.2}$$

On the other hand, the maximum likelihood estimation in the problem statement can be rewritten as:

$$\hat{\beta}_1, \hat{\beta}_0 = \arg\max_{\beta_1, \beta_0} \mathcal{L}(\{\boldsymbol{\mu}_i\}, \{\boldsymbol{\sigma}_i\}|\beta_1, \beta_0) = \arg\max_{\beta_1, \beta_0} \log(\mathcal{L}(\{\boldsymbol{\mu}_i\}, \{\boldsymbol{\sigma}_i\}|\beta_1, \beta_0))$$

$$= \arg\max_{\beta_1, \beta_0} \sum_{i=1}^{N} \frac{1}{2} \log(2\pi\boldsymbol{\sigma}_i) - \frac{(\boldsymbol{\mu}_i - \beta_1 f(\boldsymbol{v}_i) - \beta_0)^2}{2\boldsymbol{\sigma}_i^2} = \arg\min_{\beta_1, \beta_0} \sum_{i=1}^{N} \frac{(\boldsymbol{\mu}_i - \beta_1 f(\boldsymbol{v}_i) - \beta_0)^2}{2\boldsymbol{\sigma}_i^2}$$

Therefore, setting $\boldsymbol{w} = \frac{1}{2\boldsymbol{\sigma}^2}$, $\alpha = \beta_1$, $\beta = \beta_0$, $\boldsymbol{x} = f(\boldsymbol{v})$ and $\boldsymbol{y} = \boldsymbol{\mu}$ in Eq. A.2, we obtain the same optimization problem.

A.4 Model hyperparameters

We list the default parameter values, as well as the optimized set in Table A.1.

Table A.1: A list of default values used in the original work, and the optimized values used in our work.

Model	**Parameter**	**Default**	**Optimized**
Gradient Boosting	Learning rate	1.0	0.1
	№ estimators	100	200
	Maximum depth	3	2
Random Forest	№ estimators	100	50
	Maximum depth	3	5
SVM	Kernel	RBF[1]	RBF
	C	1.0	1.0

[1]RBF stands for the radial basis function kernel

Appendix B

Supplementary plots and tables

In this part of the appendix, we show any deferred plots that support the claims in the body of the report.

B.1 Sensitivity analysis

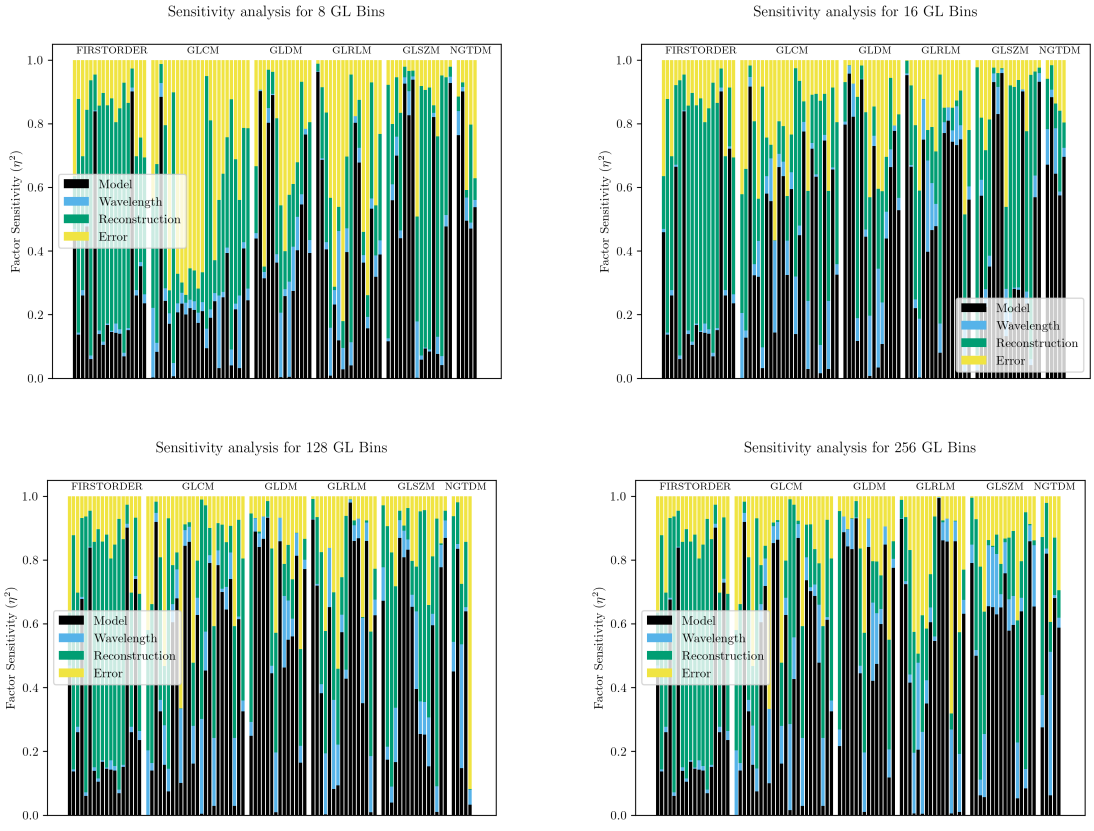


Figure B.1: Factor sensitivity when standardizing for GL bins for $N_g = 8$ (top left), $N_g = 16$ (top right), $N_g = 128$ (bottom left) and $N_g = 256$ (bottom right)

B.2 Correlation heatmaps

B.3 Statistical test results

B.3.1 Kruskal-Wallis

B.3.2 Kolmogorov-Smirnov

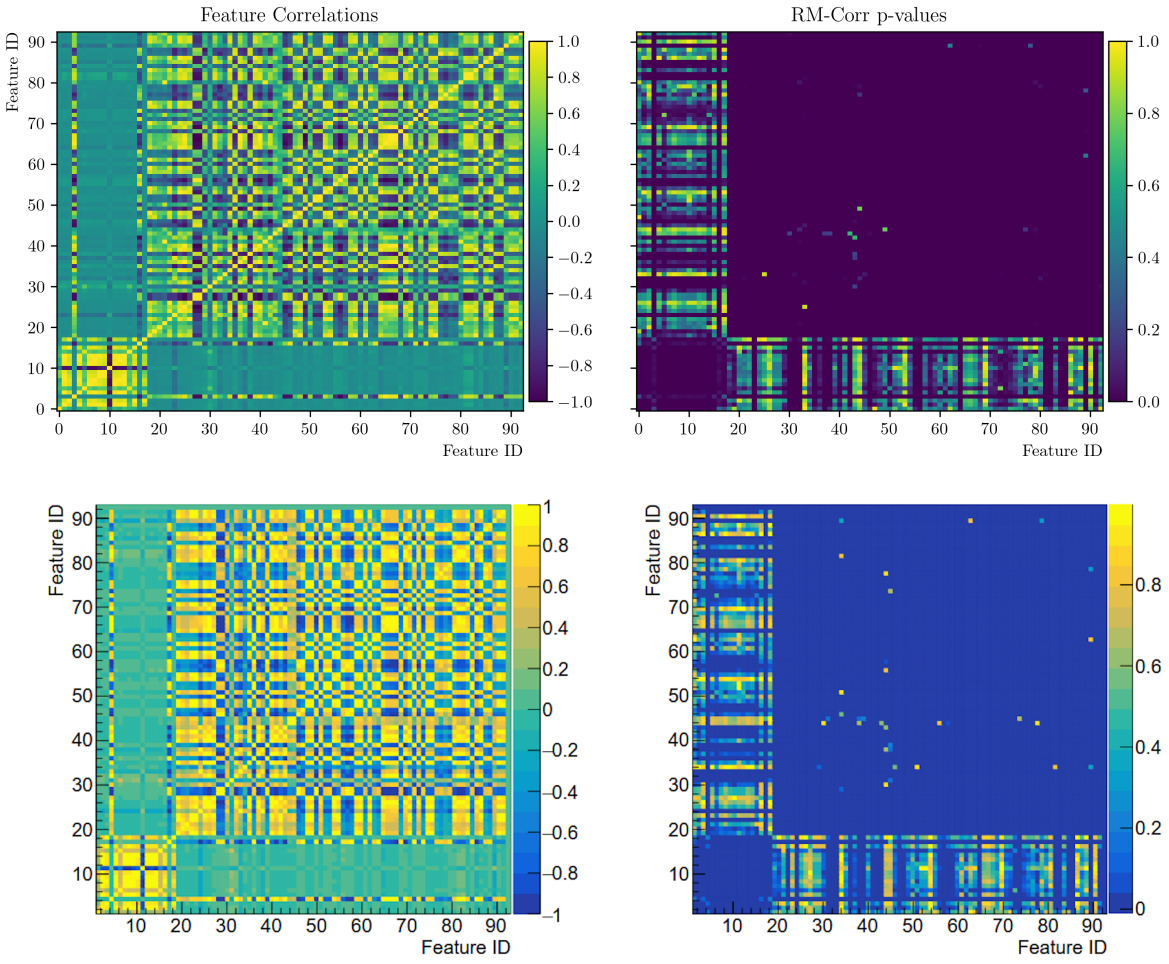


Figure B.2: Comparison between correlation measures (left) and p-values (left) for our work (top) and the original paper (bottom).

Table B.1: List of Kruskal-Wallis p-values for each feature with the adjusted p-value according to the Benjamini-Hochberg correction.

Feature Name	**P-value**	**Rank**	**Corrected value**
FO 10Percentile	7.973261872397038e-120	1	0.002688172043010753
FO 90Percentile	3.1238707745287216e-94	2	0.005376344086021506
FO RootMeanSquared	3.1238707745287216e-94	3	0.008064516129032258
FO Skewness	1.6538016539853625e-93	4	0.010752688172043012
FO Mean	5.753764688939436e-91	5	0.013440860215053764
FO Median	2.702929538589017e-80	6	0.016129032258064516
FO Variance	4.728091719579677 4e-68	7	0.01881720430107527
FO MeanAbsoluteDeviation	4.960409712484148e-66	8	0.021505376344086023
FO Maximum	1.9642395928918645e-64	9	0.024193548387096774
FO Range	4.5770847659536055e-62	10	0.026881720430107527
FO RobustMeanAbsoluteDeviation	2.887457631350044e-56	11	0.02956989247311828
FO Kurtosis	1.7602610379562294e-54	12	0.03225806451612903
FO InterquartileRange	2.09440812932595e-53	13	0.03494623655913978
GLCM Imc2	7.931124470725125e-44	14	0.03763440860215054
NGTDM Coarseness	5.569049011878999e-41	15	0.04032258064516129
GLRLM LongRunLowGrayLevelEmphasis	3.0725920172774714e-39	16	0.043010752688172046
GLCM ClusterProminence	8.660856520375989e-38	17	0.0456989247311828
GLSZM LargeAreaLowGrayLevelEmphasis	4.285559243101624e-37	18	0.04838709677419355
GLDM SmallDependenceHighGrayLevelEmphasis	4.500595861241619e-33	19	0.051075268817204304
GLSZM SmallAreaLowGrayLevelEmphasis	6.059830018769898e-31	20	0.053763440860215055
GLCM DifferenceVariance	6.636888396542592e-30	21	0.056451612903225805
FO Minimum	1.483194400097604e-28	22	0.05913978494623656
GLCM ClusterTendency	2.5933155001176547e-28	23	0.06182795698924731
GLCM SumSquares	3.233059003618089e-28	24	0.06451612903225806
GLRLM LowGrayLevelRunEmphasis	8.397360500491692e-28	25	0.06720430107526881
NGTDM Strength	3.345961174689769e-23	26	0.06989247311827956
GLCM Contrast	1.986020707984908e-22	27	0.07258064516129033
GLSZM SizeZoneNonUniformity	2.8399721865923513e-22	28	0.07526881720430108
GLRLM GrayLevelVariance	6.0620346255969675e-22	29	0.07795698924731183
GLSZM LowGrayLevelZoneEmphasis	1.1397562312549718e-21	30	0.08064516129032258
GLRLM ShortRunLowGrayLevelEmphasis	3.793038820539293e-21	31	0.08333333333333333
GLCM Imc1	1.1847978626181204e-19	32	0.08602150537634409
GLDM SmallDependenceLowGrayLevelEmphasis	1.9468010351866744e-18	33	0.08870967741935484
NGTDM Busyness	4.0453399140869854e-18	34	0.0913978494623656
GLDM GrayLevelVariance	5.224240289080862e-17	35	0.09408602150537634
GLCM ClusterShade	1.6281745732841969e-16	36	0.0967741935483871
GLRLM RunVariance	2.529892128137402e-11	37	0.09946236559139784
GLSZM GrayLevelVariance	8.701825526873265e-11	38	0.10215053763440861
GLSZM ZoneVariance	1.847258703478426e-10	39	0.10483870967741936
GLSZM LargeAreaEmphasis	2.102081043871756e-10	40	0.10752688172043011
GLSZM GrayLevelNonUniformityNormalized	1.2874712597197904e-09	41	0.11021505376344086
GLDM LowGrayLevelEmphasis	1.42319156112859e-08	42	0.11290322580645161
GLSZM GrayLevelNonUniformity	3.899015550674813e-08	43	0.11559139784946236
GLDM DependenceNonUniformityNormalized	8.914239854557365e-08	44	0.11827956989247312
GLRLM GrayLevelNonUniformity	2.317961515694604e-06	45	0.12096774193548387
GLCM MaximumProbability	2.568001332313739e-06	46	0.12365591397849462

Table B.2: List of Kruskal-Wallis p-values for each feature with the adjusted p-value according to the Benjamini-Hochberg correction (continued). Any feature below the highlighted one is rejected.

Feature Name	**P-value**	**Rank**	**Corrected value**
GLCM Idmn	9.582285968617558e-06	47	0.12634408602150538
GLSZM ZonePercentage	1.3338083069098854e-05	48	0.12903225806451613
GLSZM SmallAreaEmphasis	2.3915910358370973e-05	49	0.13172043010752688
GLRLM LongRunHighGrayLevelEmphasis	6.102382502541507e-05	50	0.13440860215053763
GLSZM SizeZoneNonUniformityNormalized	6.432358799872955e-05	51	0.13709677419354838
GLCM MCC	6.504225161402378e-05	52	0.13978494623655913
GLCM JointEnergy	8.070311598339544e-05	53	0.1424731182795699
FO Energy	9.114168192144367e-05	54	0.14516129032258066
FO TotalEnergy	9.114168192144367e-05	55	0.1478494623655914
GLRLM LongRunEmphasis	0.00032991752831851165	56	0.15053763440860216
GLRLM GrayLevelNonUniformityNormalized	0.0004067846862480478	57	0.1532258064516129
FO Uniformity	0.00079537770694731	58	0.15591397849462366
GLDM SmallDependenceEmphasis	0.0014596317164209188	59	0.1586021505376344
GLCM SumEntropy	0.0018267618048277456	60	0.16129032258064516
NGTDM Contrast	0.002815241764882808	61	0.1639784946236559
GLRLM RunEntropy	0.004587196571079836	62	0.16666666666666666
GLCM DifferenceAverage	0.007191456743549694	63	0.1693548387096774
GLDM GrayLevelNonUniformity	0.009967121161193186	64	0.17204301075268819
GLCM JointEntropy	0.016299293693311345	65	0.17473118279569894
FO Entropy	0.020273434015503567	66	0.1774193548387097
GLCM Idn	0.06878558814023546	67	0.18010752688172044
GLRLM RunLengthNonUniformity	0.09118628596118632	68	0.1827956989247312
GLDM LargeDependenceEmphasis	0.12128470388488304	69	0.18548387096774194
GLCM DifferenceEntropy	**0.2100261739190189**	**70**	**0.1881720430107527**
GLDM LargeDependenceHighGrayLevelEmphasis	0.21865079067371684	71	0.19086021505376344
GLCM JointAverage	0.24575555295271995	72	0.1935483870967742
GLCM SumAverage	0.24575555295271995	73	0.19623655913978494
GLCM Autocorrelation	0.3202608606205139	74	0.1989247311827957
NGTDM Complexity	0.3340472884770678	75	0.20161290322580644
GLRLM HighGrayLevelRunEmphasis	0.383102743862508	76	0.20430107526881722
GLCM Correlation	0.42358193966893964	77	0.20698924731182797
GLDM DependenceVariance	0.42734800119139393	78	0.20967741935483872
GLRLM RunPercentage	0.44528100009079474	79	0.21236559139784947
GLDM HighGrayLevelEmphasis	0.5213977243649499	80	0.21505376344086022
GLDM LargeDependenceLowGrayLevelEmphasis	0.541145774277377	81	0.21774193548387097
GLSZM HighGrayLevelZoneEmphasis	0.5545299136179405	82	0.22043010752688172
GLCM InverseVariance	0.6191939290348347	83	0.22311827956989247
GLCM Idm	0.6687159325316403	84	0.22580645161290322
GLSZM SmallAreaHighGrayLevelEmphasis	0.6852484582750331	85	0.22849462365591397
GLRLM ShortRunHighGrayLevelEmphasis	0.695231960304568	86	0.23118279569892472
GLDM DependenceEntropy	0.7014570770819424	87	0.23387096774193547
GLCM Id	0.7127701371701018	88	0.23655913978494625
GLSZM ZoneEntropy	0.7148775141459889	89	0.239247311827957
GLRLM ShortRunEmphasis	0.8589483223851525	90	0.24193548387096775
GLDM DependenceNonUniformity	0.8924451439812735	91	0.2446236559139785
GLRLM RunLengthNonUniformityNormalized	0.9132405794741435	92	0.24731182795698925
GLSZM LargeAreaHighGrayLevelEmphasis	0.9427269964030484	93	0.25

Table B.3: List of Kolmogorov-Smirnov p-values for each feature with the adjusted p-value according to the Benjamini-Hochberg correction.

Feature Name	**P-value**	**Rank**	**Corrected value**
FO 10Percentile	5.789231602296052e-147	1	0.002688172043010753
FO 90Percentile	4.145452767857454e-129	2	0.005376344086021506
FO RootMeanSquared	4.145452767857454e-129	3	0.008064516129032258
FO Skewness	4.468093467051294e-121	4	0.010752688172043012
FO Energy	1.983250372585267 7e-117	5	0.013440860215053764
FO Mean	1.983250372585267 7e-117	6	0.016129032258064516
FO TotalEnergy	1.983250372585267 7e-117	7	0.01881720430107527
FO Median	1.034548284931253e-103	8	0.021505376344086023
GLCM Imc2	1.8409662129816212e-90	9	0.024193548387096774
GLCM ClusterProminence	1.150854223730341e-82	10	0.026881720430107527
GLSZM LargeAreaLowGrayLevelEmphasis	5.387539898018658e-82	11	0.02956989247311828
GLDM SmallDependenceHighGrayLevelEmphasis	6.217363696737566e-78	12	0.03225806451612903
GLRLM LongRunLowGrayLevelEmphasis	1.3868637798988051e-73	13	0.03494623655913978
FO Variance	1.2615017021844322e-72	14	0.03763440860215054
GLCM SumSquares	7.381210516746658e-69	15	0.04032258064516129
GLCM ClusterTendency	1.8944795142421094e-66	16	0.043010752688172046
GLSZM SizeZoneNonUniformity	6.457790613258357e-66	17	0.0456989247311828
GLSZM SmallAreaLowGrayLevelEmphasis	9.942738549554636e-66	18	0.04838709677419355
FO Kurtosis	1.0683536288971975e-65	19	0.051075268817204304
GLSZM LowGrayLevelZoneEmphasis	5.167620468252967e-65	20	0.053763440860215055
FO MeanAbsoluteDeviation	2.744215038615665e-63	21	0.056451612903225805
NGTDM Strength	8.46126728297235e-62	22	0.05913978494623656
GLCM DifferenceVariance	4.1667028417246855e-61	23	0.06182795698924731
FO Maximum	7.237786902210203e-57	24	0.06451612903225806
GLRLM LowGrayLevelRunEmphasis	7.835877653171909e-56	25	0.06720430107526881
GLCM Contrast	3.545647586569756e-55	26	0.06989247311827956
FO Range	1.8588378097786355e-52	27	0.07258064516129033
NGTDM Busyness	3.986495619375433e-52	28	0.07526881720430108
GLRLM ShortRunLowGrayLevelEmphasis	4.2476948444952474e-52	29	0.07795698924731183
GLRLM GrayLevelVariance	1.0981267009258902e-51	30	0.08064516129032258
GLDM GrayLevelVariance	1.1963134201330609e-50	31	0.08333333333333333
FO InterquartileRange	1.7401643919205814e-50	32	0.08602150537634409
FO RobustMeanAbsoluteDeviation	7.124356318763087e-49	33	0.08870967741935484
GLSZM GrayLevelNonUniformity	1.6711408299374033e-42	34	0.0913978494623656
GLSZM ZoneVariance	5.524359285687124e-41	35	0.09408602150537634
GLSZM LargeAreaEmphasis	1.8809257956120878e-40	36	0.0967741935483871
FO Minimum	1.6802695754255972e-38	37	0.09946236559139784
GLRLM GrayLevelNonUniformityNormalized	3.2109737507343026e-38	38	0.10215053763440861
NGTDM Coarseness	9.913535239517471e-38	39	0.10483870967741936
GLSZM GrayLevelNonUniformityNormalized	2.589629831154011e-37	40	0.10752688172043011
GLDM LowGrayLevelEmphasis	5.162928853465772e-37	41	0.11021505376344086
GLDM SmallDependenceLowGrayLevelEmphasis	1.498117875702544e-34	42	0.11290322580645161
GLDM DependenceNonUniformityNormalized	1.0939953682634286e-31	43	0.11559139784946236
FO Uniformity	1.1487182699329347e-31	44	0.11827956989247312
GLSZM GrayLevelVariance	8.786763124633946e-31	45	0.12096774193548387
GLDM LargeDependenceHighGrayLevelEmphasis	5.331623276768771e-25	46	0.12365591397849462

Table B.4: List of Kolmogorov-Smirnov p-values for each feature with the adjusted p-value according to the Benjamini-Hochberg correction (continued). Any feature below the highlighted one is rejected.

Feature Name	**P-value**	**Rank**	**Corrected value**
GLDM GrayLevelNonUniformity	2.853027008083863e-23	47	0.12634408602150538
GLRLM RunVariance	3.231515124996768e-23	48	0.12903225806451613
GLCM Imc1	7.385795558829035e-23	49	0.13172043010752688
GLSZM LargeAreaHighGrayLevelEmphasis	8.291101152266545e-20	50	0.13440860215053763
GLRLM GrayLevelNonUniformity	1.6446250137422633e-19	51	0.13709677419354838
GLSZM ZoneEntropy	1.6446250137422633e-19	52	0.13978494623655913
GLSZM ZonePercentage	2.063940400434768e-19	53	0.1424731182795699
GLDM SmallDependenceEmphasis	2.3343594939277972e-15	54	0.14516129032258066
GLCM ClusterShade	4.199144545693132e-12	55	0.1478494623655914
GLRLM RunLengthNonUniformity	3.6736298454058123e-11	56	0.15053763440860216
GLDM DependenceEntropy	5.7657952759220634e-11	57	0.1532258064516129
GLDM DependenceVariance	1.4818801424996657e-10	58	0.15591397849462366
FO Entropy	1.6034315465488641e-09	59	0.1586021505376344
GLCM DifferenceAverage	1.8259387291245188e-09	60	0.16129032258064516
GLRLM LongRunEmphasis	1.8099581033527558e-08	61	0.1639784946236559
GLSZM SmallAreaEmphasis	2.671538807118824e-08	62	0.16666666666666666
GLCM InverseVariance	3.319560020595791e-08	63	0.1693548387096774
GLCM MCC	3.8339077027116534e-08	64	0.17204301075268819
GLDM LargeDependenceEmphasis	1.5676828575824443e-07	65	0.17473118279569894
GLSZM SizeZoneNonUniformityNormalized	1.6413734999691943e-07	66	0.1774193548387097
GLRLM LongRunHighGrayLevelEmphasis	1.937827716699542e-06	67	0.18010752688172044
GLCM MaximumProbability	3.307862503267385e-05	68	0.1827956989247312
GLCM Idmn	4.4616996231112356e-05	69	0.18548387096774194
NGTDM Contrast	4.717003804498426e-05	70	0.1881720430107527
GLCM DifferenceEntropy	0.00010491463173076233	71	0.19086021505376344
GLCM Idn	0.0003337153746199909	72	0.1935483870967742
GLCM JointEnergy	0.001665100840598086	73	0.19623655913978494
GLRLM RunPercentage	0.0026002749560140137	74	0.1989247311827957
GLDM DependenceNonUniformity	0.005219733234160232	75	0.20161290322580644
GLRLM RunEntropy	0.007338008224855071	76	0.20430107526881722
GLCM Correlation	0.00827489888667998	77	0.20698924731182797
GLCM SumEntropy	0.021409564071587516	78	0.20967741935483872
NGTDM Complexity	0.04104201943168825	79	0.21236559139784947
GLCM Idm	0.07031511245086458	80	0.21505376344086022
GLRLM RunLengthNonUniformityNormalized	0.07480526076128051	81	0.21774193548387097
GLRLM ShortRunEmphasis	0.08116797446993	82	0.22043010752688172
GLSZM SmallAreaHighGrayLevelEmphasis	0.0915635953173629	83	0.22311827956989247
GLCM JointEntropy	0.1135145517980647	84	0.22580645161290322
GLDM LargeDependenceLowGrayLevelEmphasis	0.16591022693893787	85	0.22849462365591397
GLCM Id	0.18124267232139002	86	0.23118279569892472
GLSZM HighGrayLevelZoneEmphasis	0.18124267232139002	87	0.23387096774193547
GLRLM HighGrayLevelRunEmphasis	**0.27065395525790253**	**88**	**0.23655913978494625**
GLCM JointAverage	0.29966131346425234	89	0.239247311827957
GLCM SumAverage	0.29966131346425234	90	0.24193548387096775
GLRLM ShortRunHighGrayLevelEmphasis	0.3993118798438525	91	0.2446236559139785
GLCM Autocorrelation	0.4021106499813008	92	0.24731182795698925
GLDM HighGrayLevelEmphasis	0.4974841182560662	93	0.25

Appendix C

Acknowledgment for the use of generative AI

We acknowledge that the use of generative AI tools in this project was strictly limited to the purposes of debugging and code documentation. These tools were utilized to enhance the clarity of the technical documentation and to assist in resolving programming-related issues. No generative AI was employed in the development of the core functionalities or intellectual content of this project.