Explainable machine learning models to assist with cancer diagnosis

Patrick Riley

A thesis submitted in partial fulfilment of the requirements of Liverpool John Moores University for the degree of Doctor of Philosophy

October 2021

Table of Contents

Abstract

This thesis proposes a package of machine learning tools to assist in the classification process of cancers through medical imaging. Enhancing the interpretability of a given machine learning method is a key focus of the work. Firstly, a “patient like me” methodology using the Fisher Information Network is created to show a robust structure of clinical data from a neural network leading to new patient cases being classified in a visual way. Next this thesis studies the partial responses of a neural network to understand how changes in values of important variables affect the contribution towards the final prediction of the classifier. This attempts to reflect clinical thinking, where changes in variables would change the clinical outcome of a decision-making process. Finally, the thesis looks at multimodality data fusion to utilise as much of the abundance of available clinical data as possible. This work looks at the effectiveness of information with an approach that includes a feature selector.

The three aspects of work have been assessed with publicly available clinical datasets to allow for clinical meaning to be ascertained. The thesis looks at potential clinical impact throughout and how the application of machine learning can be useful in a clinical setting, rather than only providing a classification output.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application or another degree of this or any other university or institute of learning.

Acknowledgements

Firstly, I must thank my supervisory team who have been fantastic to work with. Dr Sandra Ortega-Martorell, Dr Ivan Olier and Prof Paulo Lisboa, you have all been incredibly supportive and knowledgeable throughout the last three years and I cannot thank you enough. I have enjoyed our conversations both mathematical and those less so. To Sandra, your versatility and modesty have been incredibly humbling and something I will take away from this experience with a great fondness.

I'd like to thank LJMU, FET and the Mathematics department – in all its various names over the years – for both the financial and moral support which have been essential in allowing me to produce this thesis. To all the other PhD students in the department, thank you for feeding the coffee addiction… and sorry for eating more of those Christmas biscuits! Thank you to Tricia, Natasha, and Alexia for their support too.

To my friends and family who have provided copious amounts of both cake and gin over the last few years, it's my turn now! I must say hello to both Chief (Sean) and Chef (Cormac) who are amazing and to Michael for insisting on some of the terminology used in the thesis. Thank you to Mum & Dad for getting me this far, giving me quite the work ethic.

To two people who could not have been any more supportive. To my nan, June and my partner, Steven. Thank you for putting up with me and keeping me focussed on this degree. I will treasure the encouragement you've provided me with and love you both hugely. When you read this, I promise it's really done!

List of acronyms

ANOVA	Analysis of Variance
AUC	Area Under (the) Curve
BIRADS	Breast Imaging Reporting and Data System
CAD	Computer Aided Design
CBIS-DDSM	Curated Breast Imaging Subset (of the) Digital Database for Screening Mammography
CC	craniocaudal
CLAHE	Contrast Limited Adaptive Histogram Equalisation
CNN	Convolutional Neural Network
CT	Computerized Tomography
DCE	Dynamic Contrast-Enhanced (MRI)
DDSM	Digital Database for Screening Mammography
DICOM	Digital Imaging and Communications in Medicine
DWI	Diffusion Weighted Imaging
FI	Fisher Information
FIN	Fisher Information Network
GLCM	Grey Level Co-occurrence Matrix
GLN	Grey level non-uniformity
GLRLM	Grey Level Run Length Matrix
GTM	Generative Topographic Mapping
HGLRE	High grey level run emphasis
ICE	Individual Conditional Expectation
KL	Kullback-Leibler
LGLRE	Low gret level run emphasis
LIME	Local Interpretable Model-agnostic Explanations
LLE	Locally Linear Embedding
LR	Logistic Regression
LSTM	Long Short-Term Memory
MDS	Multidimensional Scaling

MLO	mediolateral oblique
MLP	Multilayer Perceptron
MRI	Magnetic Resonance Imaging
PCA	Principal Component Analysis
PET	Positron Emission Tomography
PSA	Positive Specific Antigen
RBF	Radial Basis Function
RF	Random Forest
RLN	Run length non-uniformity
ROI	Region of Interest
SOM	Self Organising Maps
SRHGLE	Short run high grey level emphasis
SS	Squared Stress
SVM	Support Vector Machines
TN	True Negative
TP	True Positive
t-SNE	t-distributed Stochastic Neighbourhood Embedding

List of figures

List of tables

1 Introduction

Breast cancer is the most frequent cancer among women, affecting 2.1 million women every year, and causing the greatest number of cancer-related deaths amongst them [1]. This type of cancer usually takes time to develop, with symptoms becoming evident later. Currently, there is not an effective way to cure late-stage breast cancer, and therefore early and accurate detection of the breast cancer tumour is critical for improving both prognosis and therefore treatment planning. This thesis focusses on breast cancer and effective diagnostic tools which use machine learning techniques to enhance the detection and classification process, while still holding explainability as a key focus throughout.

This research acknowledges that there are many machine learning applications in the area of breast cancer classification. This work looks towards enhancing the explainability of these in which limitations exist within much of the literature. Rather than conducting the classification alone, this work looks at how different concepts can be utilised in healthcare settings, where the understanding of the decision-making process is important.

Broadly this thesis can be split into three sections, for which all share the common aim for the machine to work “hand in hand” with multidisciplinary healthcare teams. These are machine learning-based tools to assist within the clinical application of breast cancer detection and classification:

- The first section looks at defining a mathematically robust structure of clinical data (e.g. patients) that can be visualised as a whole. The aim was to provide a data space where all the cases could be seen and grouped according to certain characteristics. This investigates the use of a distance-based metric focussing on the distances between observations or cases, leading to the creation of a data space that can be visualised, including the addition of new (unseen) data.
- The second section studies the effect of different variables in a dataset and how they impact the outcome of a predictive model. The aim was to expose

which variables were impacting the outcome of the model to a more significant extent.

- The third section looks at multimodality data fusion to improve the predictive capability of a classifier by appending one dataset with further information. The aim here was to have a more comprehensive view of each observation or case, using any additional information available, partly simulating how clinicians work using a varied base of evidence to support their decision.
    - A self-contained supporting chapter that also performs data fusion is presented on prostate cancer data.

1.1 Scope of the thesis

The key goal of this thesis is to provide a set of tools that assist in the classification of breast cancer lesions using mammography images. An important aspect of the presented research is the extensions of the methodologies presented and how the explainability of the models is enhanced.

Early detection is the key primary mechanism to prevent the development of breast cancer. Regular screening tests are an important aspect of the process of finding suspicious lesions pre-symptom development, and are effective in saving lives [2] as well as reducing psychological impact for patients. Mammography is a type of medical imaging that uses X-rays to capture images of the internal structures of the breast. Mammography is an effective method for breast cancer screening and to take breast imaging for breast cancer assessment with Bird et al. estimating the sensitivity of screening mammography to be between 85% and 90% [3].

A “second reader” within the decision-making process using medical imaging is a well-found concept, with “decision-aids” being something that has been considered for a long time now [4]. Double reading of mammograms has been found to be effective in improving the sensitivity of screening mammograms [5], although finding further radiologists to review these mammograms is difficult and also they may disagree [6]. Before an invasive biopsy is taken to ascertain a pathology reading – which can cause apprehension by patients – this check can be a triage

point to decide whether a biopsy would be necessary. An automated method that serves the purpose of the second reader could reduce the practical and human burden.

Furthermore, mammography is cheaper than alternatives such as MRI, with equipment and expertise already available. MRI can improve on some aspects of the process such as earlier detection but at the cost of a higher rate of false positives at higher breast densities [7]. The National Institute for Health and Care Excellence do have a stratified surveillance programme for patients in different risk groups (including breast density) on who should receive mammograms, MRI or both [8]. Breast density as a first step to infer information from mammograms can be limited in its success as the assessment of breast density is subjective and can vary across radiologists [9].

An abundance of patient data is available from healthcare records with various applications in the literature making use of this sort of data.The Gail model is prominent in the literature for assessing demographic risk factors of first-degree relatives with breast cancer [10], which has since been shown to have poor performance [11].

Machine learning in healthcare, and particularly breast cancer classification, has been proven to be successful. Various approaches towards the problem area [12]–[16] are able to discriminate between tumours and classify effectively in a manner of approaches. More recently, deep learning approaches [17]–[19] attain strong predictive outcome measures. The interpretability of machine learning algorithms varies and as machine learning algorithms continue to enter clinical practice is a known concern in the area of radiology [20]. The work in this thesis attempts to make these methods more explainable and to reflect clinical thinking in practice, for which this thesis contributes to the knowledge alongside other available applications in the area. Including the use of images alongside other auxiliary data, such as augmenting mammograms and clinical data in order to predict malignancy of cancers [17] has been proven to be successful.

The first workflow presented in the thesis uses Fisher Information Network to curate “patient like me” approaches for breast cancer diagnosis. This creates a robust representation of clinical data from a neural network classifier. In effect, the probability density estimates are used to create a visualisation of the latent space of patients from which underlying patterns and structures can be seen and understood. This approach shows a spatial representation of each patient case. It is shown that utilising extra metadata can further develop additional insights into patient groupings. The Fisher Information metric is a natural statistical measure of dissimilarity for small changes between each of the data points, according to their degree of relevance with respect to their class membership. The developed “patient like me” approach allows for mammogram data of a new patient to be projected into the learnt embedding, leading to a proposed machine learning-based triage methodology.

After considering this process of triage and similar patient groupings it was considered appropriate to understand how different variables inputted into a classifier from a dataset affect the classification and to what extent. The next workflow presented in this thesis studies the partial responses of a neural network classifier. This studies the effect of a change in each variable against its contribution to the logit to “open the black box” of a neural network. This allows for the explanation of changes in variables affecting outcomes which is important in clinical practice. These partial responses give rise to how each variable can affect the outcome of model predictions, adding insight to a machine learning-based approach to cancer classification.

Throughout the study it is noted that much data is used to gain extra clinical insight into the results, to inform and add impact to the discussion. As mentioned, the use of different modalities of data has developed in the literature over time, from risk profiling to classification. The final piece of work presented considers augmenting classifiers with various modalities of data to assess improvements in results by following a data fusion framework [21]. This work assesses both standalone and combinations of features extracted from a deep learning network trained on mammograms, hand crafted statistical texture features and lesion metadata. Of

particular importance here is how the metadata can impact the predictive capabilities of the classifier. This work attempts to reflect clinical thinking in a machine learning application, whereby the addition of further knowledge including some clinical can influence the accuracy of a decision.

It could be proposed that methods within this work could enhance and further inform screening programmes for breast cancer. In the UK, women are able to access breast cancer screening every 3 years [22] between the ages of 50 and 70 with mammograms taken initially. Furthermore, due to the COVID-19 pandemic, it is widely reported that there is a backlog of patients awaiting cancer care [23] as well as research suggesting lowering the age of beginning cancer screenings is likely to be effective [24]. This thesis provides methods that harness the power of strong machine learning methodologies alongside interpretable results that could act as a second reviewer in the process.

1.2 Aims and objectives of the presented work

The key aim of this thesis is to propose a novel representation of clinical data to assist the classification and understanding of breast cancer cases. The second aim is to investigate a selection of tools, based on machine learning methods, to assist in the classification and triage of breast cancer patient cases through mammograms. To achieve these research aims, several objectives are considered.

- Curate a robust and novel representation of clinical data which visualises the thinking of a machine learning classifier, discriminating between benign and malignant tumours. Use associated metadata to attain extra insights.
- Develop a “patient like me” approach to assess new cases within the visualisation which can lead to a triage-based system for new patient data cases.
- Propose a method that assesses the internal workings of an artificial neural network. This assesses the responses of each feature provided to the algorithm and how different values of the feature influence the prediction.
- Investigate how an abundance of available clinical data can be best utilised in a machine learning process using multimodal data fusion. As this includes

the use of medical imaging with mammograms, the branch of deep learning multimodal fusion is studied.

These aims and objectives share a common purpose; to uphold some level of explainability in the process alongside assisting multidisciplinary healthcare teams in the task of developing treatment care plans for patients, while harnessing the power of machine learning methodologies.

1.3 Contributions within the thesis

- The use of the Fisher Information Network methodology, which enhances differences and similarities between datapoints, to produce a robust representation of patients in a visualisation where similar cases are close together and dissimilar cases are represented further apart.
- The use of the FIN embedding or visualisation as a “patient like me” approach, which can be used as a triaging tool for cancer classification.
- The estimation of the projection of test cases into the trained embedding is a novel contribution, which allows the study of new cases for diagnosis and management as they arise.
- The representation of imaging data instead of tabular data in the Fisher Information Network methodology.
- The use of the partial responses of an artificial neural network to assess each feature and its impact on the final classification. The contribution studies one variable change at a time and to reflect how a variable affects the outcome of the model predictions.
- The option of using the change in value of a given feature as a tool to provide clinical context in the triaging of cancer cases using partial responses.
- The opportunity to replicate the thinking of a clinician (who will use multiple data sources to support a decision) with the application of data fusion approaches that combine imaging, statistical texture features and metadata to inform decision making. The use of statistical texture features is appreciated on a cross-disciplinary basis and form a contribution.

1.4 Thesis structure

The **next chapter** contains the literature review of the three topics presented, looking at methods to robustly represent clinical data, the effect of variables in a dataset on influencing its outcome and multimodality data fusion.

Chapter 3 of the thesis defines the dataset used in most of the work presented, the publicly available CBIS-DDSM dataset [25].

Chapters 4 and 5 define the methods and presented applications, respectively, of the Fisher Information Network methodology in the presented research. Through the development of a Multi-Layer Perceptron (MLP) and defining a robust metric, these chapters define an approach to represent the thinking of an artificial neural network leading to the "patient like me" concept described earlier. This is proposed as a triaging tool to a clinician, where a new patient case can be initially classified for further investigation. This work studies an application using statistical texture features of mammogram patches and another application using features derived from a deep learning model as to appreciate its power in the machine learning literature. This is further informed with the use of patient-level metadata.

Chapter 6 considers the use of partial responses to study how different features used within a classifier influence the prediction. Both on a practical and machine learning level, it is known that different factors will influence an outcome to different extents. This work considers how an artificial neural network responds to different statistical texture features and patient-level metadata in classifying the pathology of lesions. The contribution to the logit for differing values of important variables within the classifier are studied.

Chapters 7 and 8 look towards the use of multimodality data fusion. Combining different modalities of data allow for the impact of each data subset to be reviewed. In part, this work attempts to replicate the thinking of clinical staff, whereby various data sources contribute to the decision-making process for a given patient to varying degrees. **Chapter 7** focusses on "deep multimodality data fusion" at combinations of convolutional neural network features, statistical texture features and metadata under different data fusion structures to consider the level

of impact of each and their underlying interpretability. **Chapter 8** acts as a standalone chapter studying a prostate cancer data problem. Both imaging and metadata are leveraged here using support vector machines to discriminate between different lesion severities alongside a visualisation-informed analysis.

1.5 Publication list

The core of the work produced during this project has been published (or planned to be) as indicated below:

- **P. Riley**, I. Olier, R.G. Raidou, R. Casana-Eslava, M. Srivastava, P. Lisboa, C. Palmieri, S. Ortega-Martorell. "A novel representation of texture feature data from mammography images using Machine Learning and Visualisation to assist breast cancer diagnosis". *In preparation. To be submitted* to PLOS ONE [Within chapter 4 of thesis]
- **P. Riley**, I. Olier, R.G. Raidou, R. Casana-Eslava, M. Rea, P. Lisboa, L. Shen, C. Palmieri, S. Ortega-Martorell. "A novel visualisation to help characterise breast cancer patients using deep learning and Fisher Information Networks on mammograms". *Submitted to Scientific Reports* 2021. [Within chapter 4 of thesis]
- **P. Riley**, I. Olier, M. Rea, P. Lisboa, S. Ortega-Martorell. "A voting ensemble method to assist the diagnosis of prostate cancer using multiparametric MRI". 13th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (WSOM+) 2019. Barcelona, Spain. June 2019. [Chapter 8 of thesis]
- M. Srivastava, I. Olier, **P. Riley**, P.J.G Lisboa, S. Ortega-Martorell. "Classifying and grouping mammography images into communities using Fisher information networks to assist the diagnosis of breast cancer". 13th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (WSOM+) 2019. Barcelona, Spain. June 2019.
- **P. Riley**, I. Olier, P. Lisboa, S. Ortega-Martorell. "Using partial responses to enhance the interpretability of a breast cancer classifier". *In preparation. To*

be submitted to International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO) 2022. [Chapter 6 of thesis]

Additional publication during this thesis:

- S. Ortega-Martorell, A.P. Candiota, R. Thomson, **P. Riley**, M. Julia-Sape, I. Olier. "Embedding MRI information into MRSI data source extraction improves brain tumour delineation in animal models". PLOS ONE 2019, 14(8):e0220809. DOI: 10.1371/journal.pone.0220809

2 Literature review

This chapter reviews the most relevant literature around the three aspects studied in the thesis. The first section reviews both supervised and unsupervised metric learning methodologies to define a robust representation of clinical data. Furthermore, an associated method using community structure is presented. The following section reviews methods that assess the effect of variables within a "black box" machine learning model. The final section reviews multimodality data fusion, more specifically those that utilise deep learning or features derived from deep learning applications in healthcare settings.

2.1 Metric learning to define a robust representation of clinical data

This section relates to the creation of a mathematical representation of a machine learning classifier through the Fisher Information Network – using distances as a measure of similarity. This seems instinctive, where some given representation keeps similar instances closer together. However, any measure like this must be well-defined.

A metric [26] over a vector space X is a function with mapping $d: X \times X \rightarrow \mathbb{R}$, such that the following properties hold for all of $x_i, x_j, x_k \in X$:

1. Non-negativity: $d(x_i, x_j) \geq 0$
2. Symmetry: $d(x_i, x_j) = d(x_j, x_i)$
3. Triangle inequality: $d(x_i, x_k) \leq d(x_i, x_j) + d(x_j, x_k)$
4. Discernability: $d(x_i, x_j) = 0 \leftrightarrow x_i = x_j$

Where the discernability condition is not met, the metric is known as a *pseudometric*. This can often be the case and so throughout the thesis, *metric* is the term used for both cases.

2.1.1 Unsupervised metric learning

Firstly, unsupervised methods for metric learning are considered. These do not utilise anything more than the dataset itself and so the labelling of each case is disregarded. Therefore, the data features and its (co)variance are all that is considered to assess similarity for instance. These methods are linked with

dimensionality reduction methods, another key area in machine learning [27]. These methods learn a minimal, lower-dimensional representation of the original input dataset through the retention of the geometric relationships between the data instances in the original input space under some metric. This lower-dimensional representation, which still holds a given level of variance between the points inherently means the distances become more useful than previously.

Principal component analysis (PCA) [28], [29] is a commonly-used linear data projection method. It aims to reduce the dimensionality of the dataset with many interrelated variables while retaining as much as possible of the variation present in the dataset. The transformed output is a linear combination of independent, uncorrelated variables - principal components - as coordinates within the new feature space. The first principal components hold the most variance – the key aspect of the data is kept discerning the differences throughout the dataset.

The technique of PCA is described in [30]. For a dataset, $\boldsymbol{X}$ of size N x L (where N is features and L is number of cases), PCA first calculates the covariance matrix, **S**, where $\overline{\boldsymbol{X}}$ is the standardised dataset.

$$\boldsymbol{S} = \frac{\overline{\boldsymbol{X}}\overline{\boldsymbol{X}}^{\boldsymbol{T}}}{L-1}$$

Equation 2-1

The principal components of the data are provided through the eigen-decomposition of the covariance matrix, where $\boldsymbol{U} = (\boldsymbol{u_1}, \boldsymbol{u_2}, \dots, \boldsymbol{u_N})$ is the N x N matrix of eigenvectors u_i as columns and Λ is a diagonal matrix with elements $\Lambda_{\mathrm{ii}} = \lambda_i$ corresponding to the eigenvalues of $\boldsymbol{u}_i$. The first principal component is the eigenvector with the largest eigenvalue, further, it holds the most variance, descending as per the second eigenvector.

$$\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Lambda}\mathbf{U}^{-1}$$

Equation 2-2

After calculating the eigen-decomposition, the transformation of the data is obtained where U_M is a matrix that contains the M eigenvectors included in the

transformation columns. The variance of the data – what is of interest – is preserved through holding the first set of eigenvectors with the largest eigenvalues.

$$\boldsymbol{Y} = \boldsymbol{U}_M^T \boldsymbol{X}$$

Equation 2-3

Multidimensional scaling (MDS) [31] hosts a group of algorithms not based on linear projections. These algorithms begin with a matrix of pairwise dissimilarities between every point for which MDS produces a representation of this matrix in a space in which the distances between each point approximate as closely as possible the dissimilarities between the corresponding points in the original data matrix.

MDS takes a pairwise disparities matrix, $\hat{d}$, calculated as a mapping of the dissimilarities, d. A configuration of points in the Euclidean space is calculated that minimise a cost function, such as squared stress:

$$SS = \sum_r \sum_s \sum_i (d_{rs,i}^2 - \widehat{d^2}_{rs,i})$$

Equation 2-4

t-distributed stochastic neighbour embedding (t-SNE) [32] aims to represent each object from the high-dimensional space by a point in a two- or three-dimensional scatter plot, and to arrange the points in a way that similar objects are modelled by nearby points, whilst dissimilar objects are modelled by distant points with high probability. The t-SNE algorithm includes two main stages:

Firstly, the algorithm constructs a probability distribution over pairs of high-dimensional objects, which are proportional to the similarity of objects (similar objects have a high probability of being selected whilst dissimilar points have a small probability of being selected).

Secondly, the algorithm defines a similar probability distribution over the points in the low-dimensional map, and it minimizes the non-symmetric Kullback-Leibler (KL) divergence between the two distributions with respect to the locations of the points in the map. It gives a numerical representation for the deviation between

two model distributions. Say Q is a model distribution and P is a true distribution, $KL(P||Q)$ is defined as:

$$KL(P||Q) = \sum_i P_i \log\left(\frac{P_i}{Q_i}\right)$$

Equation 2-5

It can be understood as an approximate measure of how much information is lost between the model and true distribution. By viewing this as a measure of similarity, optimising the target distribution, minimising the KL divergence to the true distribution is the link to t-SNE.

A heavy-tailed Student-t distribution is used to measure similarities between low-dimensional points to allow dissimilar objects to be modelled far apart in the map. The minimization of the KL divergence with respect to the points in the lower dimensional space is performed using gradient descent. This results in a lower-dimensionality map that reflects the similarities between the high-dimensional inputs.

Locally linear embedding (LLE) [33] characterises global structure through an analysis of the local neighbourhoods keeping all mapping local. Coefficients that best reconstruct each data point from its neighbours are calculated. These are arranged to be invariant to adjustments (translations, rotations etc) of that data point and its neighbours to ensure they characterise the local geometrical properties of the neighbourhood.

The LLE algorithm then maps the high-dimensional data down to a lower-dimensional space while preserving these learned neighbourhood coefficients. The transformation of a linear local neighbourhood can be achieved through adjustments to preserve the angles formed between the data points and their neighbours. As the weights are invariant to these transformations the same weight values will reconstruct the data points in the lower-dimensional space as in the higher-dimensional space.

Generative topographic mapping (GTM) [34] is a nonlinear latent variable model of the manifold learning family, with sound foundations in probability theory. It

performs simultaneous clustering and visualization of the observed data through a nonlinear and topology-preserving mapping from a visualization latent space (with being usually 1 or 2 for visualization purposes) onto a manifold embedded in a multi-dimensional space, where the observed data reside. The mapping that generates the manifold is carried out through a generalized additive regression function:

$$y = W\varphi(u)$$

Equation 2-6

where $y \in D, u \in, W$ is the matrix that generates the mapping, and φ is a vector with the images of S basis functions φs. To achieve computational tractability, the prior distribution of u in latent space is constrained to form a uniform discrete grid of M centres, analogous to the layout of the Self-Organizing Map (SOM) units, in the form of a sum of delta functions:

$$p(\boldsymbol{u}) = \frac{1}{K}\delta(\boldsymbol{u} - \boldsymbol{u_k})$$

Equation 2-7

GTM typically uses a set of radial basis functions to map the results of the unsupervised analysis.

2.1.2 Supervised metric learning

Supervised metric learning does consider the labelling of cases. Similarity can be calculated therefore with respect to the target variable. These methods can lead to classification processes as demonstrated within this thesis.

Nearest neighbour applications are very common and well-known throughout the metric learning literature and are of importance to note [35]. This is a memory-based method where no model is fit, in which the distance metric heavily influences the definition of the neighbourhood for determining the classification of cases. Unclassified points are assigned a label based on the nearest set of classified points [36] through a nonparametric prediction of a target function, $f(x)$, at a given point x_0 in the input space based on known values of the function in the surrounding area, where $C(x_0)$ is the set of neighbours of x_0 used for the prediction (see

Equation 2-8). It is assumed that $f(x)$ is a smooth function that is approximately constant within the neighbourhood.

$$f(x_0) = \frac{1}{|C(x_0)|} \sum_{x_i \in C(x_0)} f(x_i)$$

Equation 2-8

Considering a classification setting with data points, $X = \{x_1, \dots, x_L\}, x_i \in \mathrm{R}^{\mathrm{N}}$, with known class labels $Y = \{y_1, \dots, y_L\}, y_i \in \{1, \dots, J\}$. Although conditional probabilities would be useful for the target function, they are unlikely to be known and so an estimation can come from the cases $\{x_i, y_i\}$. Following the principle in Equation 2-8, the probability $p(c_j(x) = 1|x_0)$ can be estimated as shown in Equation 2-9.

$$\hat{p}(c_j|x_0) = \frac{1}{K} \sum_{x_i \in C(x_0)} c_j(x_i)$$

Equation 2-9

To classify a new point, the K nearest points from the training data set are identified and assign the new point to the class having the largest number of representatives amongst this set. A "true" nearest neighbour representation is where K = 1, or the single nearest neighbour to a given point is identified.

Large margin nearest neighbour classification [37] learns a Mahalanobis distance metric from data that attempt to rearrange the data space so that the K-nearest neighbours of each point always belong to the same class, pushing data points from other classes further away. This is through minimising a cost function to find a transformation matrix that achieves this separation and is framed as an optimisation problem that finds a global minimum.

Kernels map data points, $\boldsymbol{x}$, from the original input space to a new, higher dimensional space, $\emptyset(\boldsymbol{x})$, where a new linear procedure is then applied. The new higher dimensional space can perform linear methods that have a non-linear effect in the original input space. Kernel methods are not mapped explicitly using inner products between mapped points, obtained from the kernel function that defines them (or the kernel trick), as in Equation 2-10. These kernel functions – linear,

polynomial, radial and others – can represent measures of similarity between pairs of points within a feature space.

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \emptyset(\boldsymbol{x}_i).\emptyset(\boldsymbol{x}_j)$$

Equation 2-10

In relation to similarity measures, kernels are related to distance measures. By performing the kernel trick where $\boldsymbol{x} \in \mathbb{R}^N \rightarrow \emptyset(\boldsymbol{x}) \in \mathbb{R}^F, F \gg N$ is that the original input space is mapped into a new region, $\mathcal{M}$ in a high-dimensional feature space. The mapping function is what determines the geometric structure of $\mathcal{M}$. Satisfying some assumptions, $\mathcal{M}$ will be an N-dimensional differentiable manifold in the feature space, containing the image of $\mathbb{R}^N$.

This mapping, $\emptyset(\boldsymbol{x})$, also includes a metric in the input space. This metric relates to measuring distances between pairs of images $\emptyset(\boldsymbol{x_i})$ and $\emptyset(\boldsymbol{x_j})$ in the feature space along the surface $\mathcal{M}$:

$$d_{\emptyset}(x_i, x_j) = d_{\mathcal{M}}(\emptyset(\boldsymbol{x_i}), \emptyset(\boldsymbol{x_j}))$$

Equation 2-11

$\mathcal{M}$ will usually be a curved surface known as a Riemannian manifold, with distances calculated within them worked out locally using a Riemannian tensor, $\boldsymbol{G}(\boldsymbol{x})$:

$$d_{\emptyset}(\boldsymbol{x}, \boldsymbol{x} + d\boldsymbol{x})^2 = d\boldsymbol{x}^T \boldsymbol{G}(\boldsymbol{x}) d\boldsymbol{x}$$

Equation 2-12

$\boldsymbol{G}(\boldsymbol{x})$ is a symmetric positive definite matrix that determines the metric. To express it in terms of the kernel function, consider the length of an infinitely short segment, $d\boldsymbol{z}$ in $\mathcal{M}$. Taking $\boldsymbol{z} = \emptyset(\boldsymbol{x})$ and $\boldsymbol{z} + d\boldsymbol{z} = \emptyset(\boldsymbol{x} + d\boldsymbol{x})$, the mapped segment and its length are as follows, using the first-order Taylor approximation of $\emptyset(\boldsymbol{x} + d\boldsymbol{x})$.

$$dz = \emptyset(\boldsymbol{x} + d\boldsymbol{x}) - \emptyset(\boldsymbol{x}) \approx \nabla_{\mathbf{x}} \emptyset(x) d\boldsymbol{x}$$

Equation 2-13

$$d_{\emptyset}(\boldsymbol{x}, \boldsymbol{x} + d\boldsymbol{x})^2 = \|d\boldsymbol{z}\|_2^2 = d\boldsymbol{x}^T \nabla_{\mathrm{x}} \emptyset(\boldsymbol{x})^T \nabla_{\mathrm{x}} \emptyset(\boldsymbol{x}) d\boldsymbol{x}$$

Equation 2-14

Utilising the definition of the kernel function, $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \emptyset(\boldsymbol{x}_i).\emptyset(\boldsymbol{x}_j)$, the expression of the Riemannian metric in terms of the kernel function used, is attained:

$$d_{\emptyset}(\boldsymbol{x}, \boldsymbol{x} + d\boldsymbol{x})^2 = d\boldsymbol{x}^T \nabla_{\mathrm{x}} \nabla_{\mathrm{y}} K(\boldsymbol{x}, \boldsymbol{y})|_{\boldsymbol{y}=\boldsymbol{x}} d\boldsymbol{x}$$

Equation 2-15

To calculate the distances between points that are not adjacent in $\mathcal{M}$, Equation 2-12 is transformed into a path integral:

$$d_{\emptyset}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \left| \int_{t_i}^{t_j} \sqrt{\boldsymbol{x}(t)^T \boldsymbol{G}(\boldsymbol{x}(\boldsymbol{t}))\boldsymbol{x}(t)}\ dt \right|$$

Equation 2-16

The **Fisher kernel** [38] was developed to combine generative and discriminative models in classification – it measures the similarity of two points with reference to a generative statistical model, $p(x|\theta)$. The generative model produces a manifold where the Riemannian metric tensor is the Fisher information matrix ($\boldsymbol{G}(\boldsymbol{\theta})$), where E_x is the expectation with respect to $p(x|\theta)$

$$d(\boldsymbol{\theta}, \boldsymbol{\theta} + d\boldsymbol{\theta})^2 = d\boldsymbol{\theta}^T \boldsymbol{G}(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

Equation 2-17

$$\boldsymbol{G}(\boldsymbol{\theta}) = E_x(\nabla_{\theta} \log\, (p(\boldsymbol{x}|\boldsymbol{\theta}) \nabla_{\theta} \log p(\boldsymbol{x}|\boldsymbol{\theta})^T)$$

Equation 2-18

The distance between two points, $\boldsymbol{\theta}$ and $\boldsymbol{\theta} + d\boldsymbol{\theta}$ within this metric corresponds to the distance between the two corresponding densities, $p(\boldsymbol{x}|\boldsymbol{\theta})$ and $p(\boldsymbol{x}|\boldsymbol{\theta} + d\boldsymbol{\theta})$ along the manifold. This measures how different they are in terms of the expected variation in the log-likelihood of $\boldsymbol{x}$**.** The metric is Riemannian and not Euclidean and so the expected variation of $\log p(\boldsymbol{x})$ caused by a distortion in θ can be different, depending on the location of the space in which it is measured.

The Fisher kernel is therefore defined, of two points in the input space given a generative model $p(\boldsymbol{x}|\boldsymbol{\theta})$:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{n}(\theta, \boldsymbol{x}_i)^T . \boldsymbol{G}(\theta) . \boldsymbol{n}(\theta, \boldsymbol{x}_j)$$

Equation 2-19

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \nabla_\theta \log p(\boldsymbol{x}|\theta)^T . \boldsymbol{G}(\theta)^{-1} . \nabla_\theta \log p(\boldsymbol{x}|\theta)$$

Equation 2-20

$\boldsymbol{n}(\theta, \boldsymbol{x})$ is the natural gradient, used to find the θ direction of the steepest ascent of the log-likelihood at point $p(\boldsymbol{x}|\boldsymbol{\theta})$ of the manifold:

$$\boldsymbol{n}(\theta, \boldsymbol{x}) = \boldsymbol{G}(\theta)^{-1} \nabla_{\text{theta}} \log p(\boldsymbol{x}|\theta)$$

Equation 2-21

2.1.2.1 Fisher Information metric in the input space

The Fisher Information (FI) metric is informed about the generative probabilities of the data and can only assess the importance of directions in the space of the parameters. A connection between the FI metric and Kullback-Leibler divergence exists to reinforce the role of the FI metric as a measure of the difference between adjacent probability distributions on the manifold:

$$I_{KL}\big(p(\boldsymbol{x}|\boldsymbol{\theta}), p(\boldsymbol{x}|\boldsymbol{\theta} + d\boldsymbol{\theta})\big) = d\boldsymbol{\theta}^T \boldsymbol{G}(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

$$where\ I_{KL}\big(p(\boldsymbol{x}|\boldsymbol{\theta}), p(\boldsymbol{x}|\boldsymbol{\theta} + d\boldsymbol{\theta})\big) = -\int \log\left(\frac{p(\boldsymbol{x}|\boldsymbol{\theta} + d\boldsymbol{\theta})}{p(\boldsymbol{x}|\boldsymbol{\theta})}\right) p(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x}$$

Traditionally [38], [39], the FI metric has been defined within the space parameters based on generative models, $p(\boldsymbol{x}|\boldsymbol{\theta})$. Based on [40], [41], it is possible to amend the approach to apply the Fisher metric on discriminative classification models, $p(y|\boldsymbol{x})$. Here, the metric measures parameter distortions with reference to the input space, $\boldsymbol{x}$, instead of θ.

2.1.3 Community structure example – HeC CaseReasoner

Other concepts that exist throughout the literature include those that use community structures. Although these are not studied in the thesis, they are an important application within the literature. An interesting application using community structures for children's healthcare is described.

A more pragmatic approach is shown through the Health-e-Child CaseReasoner project [42], [43] curated with relative neighbourhood graphs. These are adaptive to various distance metrics and can represent patient groupings well leading to much room for adjustment depending on the level of complexity required. This looked at techniques for learning discriminative distance functions being made available to clinicians. Cases are treated as vertices with two similar (close together) cases connected with an edge and are considered to be closest together with respect to a given distance function [44]. On a practical level, this application is adaptive allowing for greater clinical impact, where each case can be coloured based on a nominated attribute which can lead to nearest neighbour classification.

Firstly, the graphs are clustered appropriately, using both the Girvan and Newman's community structure algorithm [45] and a top-down induction of a semantic clustering tree to provide every cluster with a semantic description for clinical inspection (this was designed for CaseReasoner). The community structure algorithm separates networks into communities by iteratively removing edges with the largest number of shortest paths between them. These edges within a network are removed until an acceptable community split is found.

Two methods are reviewed in this study for the learning process to be optimal using distances, arguing that it is easier than sharing a black-box model. The first is to learn equivalence constraints, where pairs of cases and their labels are considered to assess whether they belong to the same class. The original feature space is transformed into a product or difference space allowing for any machine learning technique to be used to learn in the new space. This is an intuitive method as a natural input for optimal distance function learning and can map well to clinical understanding.

The random forest distance function considers the proportion of trees where two cases appear in the same leaves can be used as a measure of similarity between them [46]. For a given forest, f, the similarity between two cases, x_1 and x_2 can be calculated; each case is propagated through all K trees within the forest f with their associated terminal positions z in each of the trees (z_1=(z_{11},…,z_{1K}) for x_1, similarly z_2 for x_2) are noted. The similarity between the two instances is shown in Equation

2-22 (I is the indicator function). This measure can be used for a variety of tasks related to the classification problem including nearest neighbour classification.

$$S(x_1, x_2) = \frac{1}{K} \sum I(z_{1i} = z_{2i})$$

Equation 2-22

This approach holds inherent explainability. The points are well defined as cases with links drawn as connections to different cases. Groupings can be detected throughout. The application can easily visualise and filter in/out different selections and is easy to read. As a decision-support mechanism, a considered hypothesis can be assessed visually by a subject-matter expert to accept or reject the original thinking. This allows for a reflection on model explainability.

The presented work on the Fisher Information Network builds on the described approach but using the Fisher metric. Patient cases are mapped into an embedding for which patients with similar characteristics are intended to land closer together than others. This will use what has been inherently learned from an MLP classifier. Test cases will be projected into the embedding using MDS and similarity measures.

2.2 Rule-based models vs the "black box" – the trade-off of interpretability

Within the machine learning literature there are models that are inherently interpretable and other "black box" methods. Models that are harder to understand tend to perform better than those that are more interpretable for various reasons, including more adjustable parameters allowing for more complexity to be assessed. This section of the literature review considers this trade-off with inherently interpretable models and the move towards "black box" methods and methods to try and understand their inner workings.

Generalised Additive Models (GAMs) [47] were designed to be interpretable. Useful in lower-dimensional settings, they can utilise non-linear aspects in one or more variables. The model can be estimated using the backfitting procedure, as they are a linear combination of several univariate functions. Sparse Additive Models (SAMs) [48] extend the functional analysis of variance (ANOVA) model by adding l_1 regularisation alongside other constrains on the model parameters allowing for a

solution with convex optimisation. This method also identifies the best subset of variables to the problem at hand.

Traditional machine learning methods including logistic regression (LR) attain good results and are interpretable [49], [50]. In [49] for example, an analysis using LR to classify the presence of cancer attained strong predictive performance as well as highlighting features that were significant in predicting breast cancer. This included patient and lesion metadata but not age. This enables an informed discussion both at the machine learning and clinical level.

Logistic regression as a classifier uses independent variables, X_i that are related to the dependent variable, Y, leading to the calculation of the logit, where $p = \Pr(Y = 1)$ and $Logit(p) = \ln(\frac{p}{1-p})$.

$$Logit(p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

Equation 2-23

p can be calculated by taking the inverse of the $Logit(p)$

$$p = \frac{exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}{1 + exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}$$

Equation 2-24

The β coefficients can identify the average effect on the outcome of a single unit change in the variable, X, where holding all other predictors fixed. This can lead to the odds ratio to be calculated for a given variable, to interpret the effect of a given variable on the dependent variable and is estimated by $\exp(\beta)$.

Support Vector Machines (SVMs) [51] attempts to find a plane that separates classes within a feature space. Where data is not linearly separable, SVM methods utilise kernel methods to transform the data into a feature space where they would become linearly separable. Particularly in healthcare settings as per the work presented in this thesis, interpretability can begin to suffer compared to a method like logistic regression as data is transformed into a new feature space.

SVM classifiers are defined as shown in Equation 2-25, where $K(x_i, x_i')$ is the kernel function which allows for non-linear separation boundaries.

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x_i, x_i')$$

Equation 2-25

A common kernel function for the feature space transform is the radial kernel:

$$\exp\left(-\gamma \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2\right)$$

Equation 2-26

Belle et al [52] explored a method to explain SVMs in a way that interpretations for model-based decisions are able to be provided. Nomograms are derived in a similar way as it is possible for logistic regression. A Taylor expansion is applied to the Gaussian kernels and are reshaped by separately summing univariate and bivariate terms. To select the best subset of features, an iterative application of the kernel trick to a re-weighted objective function with l_1 regularisation is performed.

Rule-based methods are built with inherent interpretability. Decision trees [53] split data according to different cut-offs within features as much as required. The splits throughout the process are decided based on reducing a value, usually a misclassification rate in classification as much as possible. This method is very clear and mirrors human-decision making processes, as the splits are usually based on the value of a variable being larger or smaller than the split point or meeting the criterion for a categorical variable. Further, it is possible to calculate the importance of a given feature by summing how much the misclassification rate or error has reduced by the splits of a given variable. The more it is reduced, the more significant that given variable is in the process. However, decision trees tend to fall behind in predictive capability compared to more powerful machine learning methods.

Neural networks and in particular deep learning methods are more complex machine learning methods. They learn through adapting the weights of their synaptic connections. They can be as shallow or deep as required. Interpreting these is of interest throughout the literature. Feature visualisation [54] attempts to

show what features have been learned by the neural network inherently[1]. For this concept, the input that maximises the activation of a unit within the neural network leads to unmasking a learned feature. By "unit" this work means neuron, channel, layer, or class probability neuron. It is framed as an optimisation problem, where the network is trained a new image that maximises the activation of the unit (say a single neuron):

$$img* = \arg\max_{img} h_{n,x,y,z}(img)$$

Equation 2-27

The function, h, is the activation of a neuron, img the input of the network and x and y the spatial position of the neuron, n the layer and z the channel index. For a channel z within layer n (where all neurons in the channel are equally weighted)

$$img* = \arg\max_{img} \sum_{x,y} h_{n,x,y,z}(img)$$

Equation 2-28

Saliency maps [55] are for understanding the workings of image classification with deep (convolutional) neural networks. They highlight pixels within an image that were relevant for an image to be classified in a certain way by a convolutional neural network. The gradient of the loss function for the class of interest with respect to the input pixels is calculated, providing a map which is of the size of the input features providing positive or negative values.

This is achieved by forming an optimisation problem. For a linear score model for the class, c where the image I is represented in the one-dimensional form and w_c and b_c are the weights and biases of the model respectively:

$$S_c(I) = w_c^T I + b_c$$

Equation 2-29

The value of w defines the importance of the pixels within the image, for the class c.

[1] For the purposes of the description, images will be used but this could also be used with text or more regular tabular data.

Considering a convolutional neural network (CNN), the class score $S_c(I)$ will be a non-linear function of I – given an image I_0 it is possible to approximate $S_c(I)$ with a linear function in the neighbourhood of I_0 by computing the first-order Taylor expansion, where w is the derivative of S_c w.r.t the image I at the point I_0

$$S_c(I) = w^T I + b$$

Equation 2-30

$$w = \frac{\partial S_c}{\partial I}\bigg|_{I_0}$$

Equation 2-31

There have been extensions to saliency maps including Grad-CAM [56], DeconvNet [57] and SmoothGrad [58], all as some adjustments to the described saliency maps.

Interestingly there is debate around the need to study methods for interpreting black-box models in the first place, with authors such as Rudin [59] arguing that models should be designed to be interpretable in the first place rather than attempting to interpret models that may not be built for the purpose.

There are methods that are not particularly associated with a named model. Partial dependencies [60] show how one or two features affect the predicted outcome of a machine learning model. A partial dependence plot can show the relationship between the variable in question and the target variable to see whether it is linear or monotonic, for example. These are intuitive and easy to interpret although can only calculate these interactions for up to two features.

Individual conditional expectation (ICE) plots [61] extend on partial dependencies by visualising any supervised machine learning algorithm's estimations. They show the functional relationship between the predicted response and the feature, for individual observations. In effect, they show how a given case prediction can change when a feature changes. More formally, for each instance in $\{(x_s^i, x_c^i)\}_{i=1}^N$ the curve $\widehat{f_s^i}$ is plotted against x_s^i while x_c^i remains fixed. In the work presented in this thesis, partial responses show how a change in value of a given feature affects

the contribution of that variable to the outcome of an MLP, rather than changes in each case as per ICE plots.

Local interpretable model-agnostic explanations (LIME) [62] is a method that explains the predictions of any classifier through learning local surrogate models. These are trained to approximate what the underlying model predicted. In effect for a case that an explanation of the black box prediction is required, the dataset is perturbed. The black box predictions for these new points are taken, and the new samples weighted based on their proximity to the case being studied. A new, interpretable model (such as a decision tree) is trained on these new data. The local model is now interpreted.

The presented workflow on partial responses of a neural network is an attempt to open the black box of an MLP. The responses of each individual feature throughout the MLP are assessed to ascertain the strongest changes in the predictive outcome and assess a decision-making process in a clinical setting. This work looks to provide more context to the decision-making process which can sometimes be limited in the literature. The presented methodology considers how a minimal set of variables can be formed to deliver the information and considers how robust the explnations are in the given setting.

2.3 Deep learning and multimodal data fusion

Multimodal data fusion “aims to integrate the data of different distributions, sources, and types into a global space in which both inter-modality and cross-modality can be represented in a uniform manner.” [63] Through considering different data modalities within different classifiers this research area considers more succinct ways of combining other types of data to classify in the same way.

Importantly the work presented in this thesis attempts to reflect a decision-making process in modern medicine, which relies on using data from different sources to influence the outcome. It is known that the fusion of different classifiers can be useful where they are different [64] or where different sets of features are used [65].

The inclusion of deep learning combines the strength of deep learning models alongside the utilisation of other data modalities to make an effective prediction. The process is the same however in some way, deep learning is included. This is usually where features or predictions are extracted using a deep neural network, such as a convolutional neural network. The application of deep multimodal data fusion using medical imaging alongside other data modalities is considered within the thesis. A recent systematic review [21] categorises three types of deep multimodal data fusion and these are followed for the purposes of the thesis.

Early Fusion [63] joins different modalities into a single feature vector before applying it to a machine learning classifier. Either original data features or extracted features (say from extracting texture features) are applicable here.

An example from Thung et al [66] concatenated two imaging modalities, PET and MRI, patient data (age, gender, education) and genetic data for the diagnosis of Alzheimer's disease. This was conducted using a deep neural network, with both modality-specific layers and task-specific layers as a form of multi-task learning. The inclusion of all available clinical data attained the highest accuracy in this process at 63.6% accuracy, while MRI alone attained only 58%.

An et al. [67] used optical coherence tomography features, ocular parameters from laser speckle flowgraphy features and patient data for the classification of glaucoma cases. A feature selection process was implemented to improve interpretability for which the importance of variables was quantified at the end of the process, which included data from all data modalities. Although no comparison to standalone classification is provided, the fusion accuracy was 87.8%.

Tong et al. [68] performed early fusion on histology images from the ICIAR 2018 Grand Challenge[2]. They concatenate CNN features from three different networks before making a final prediction, as a 5-class classifier. All fusion models tested outperformed single modality models, the best performing attaining an accuracy of

[2] Both "early" and "late" fusion are referred to in a different manner to the adopted definitions within this thesis. Features are concatenated, not predictions and so falls under the "Early Fusion" definition.

81.5% where features were combined before the final fully-connected layers in the network, using AlexNet to extract the features.

Tan et al. [69] used feature-level fusion with various modalities of CT scan and statistical texture features to discriminate between levels of severity of lung cancer. Comparisons were made between a standalone CNN model using CT scans, the CNN model alongside knowledge from the scans which involve information from the specific patch within the image and finally a model that includes both aforementioned information alongside extracted statistical texture features. These are included in the model before any classification is made – so extracted features and statistical texture features – and so is denoted as early fusion by this thesis' adopted definitions. The inclusion of further information improves the predictive capability throughout, where the standalone model attains an area under the curve (AUC) of 0.89 the model with all information available attains an AUC of 0.943.

Early fusion using multi-sequence MRI is used in work by Feng et al. [70]. In this work, both diffusion-weighted MRI imaging (DWI) and various parameters of dynamic contract enhanced MRI imaging (DCE) are run through a feature extractor CNN separately, with these outputted features combined for a classification process to classify between benign and malignant breast cancer. This work used domain knowledge to assist in the feature learning process, utilising and extracting relevant morphological features to ensure the best usability of the model. An LSTM model [71] is used to combine the extracted features as this model is proven to work well with sequential data (in this work the MRI images display the same data in a different manner). Results are compared across standalone DCE classification, standalone DWI classification and using the proposed feature fusion model. The best results are found with the combination of all data with an accuracy of 85% at the patient level, compared to 80% for both standalone models.

Joint fusion considers joining learned feature representations from intermediate layers of neural networks. Features from other data modalities are inputted at the end. In this type of fusion, the loss is backpropagated to the feature extracting neural networks to represent data better.

Yala et al. [72] performed a larger-scale retrospective study using deep learning to develop a breast cancer risk model. Mammograms were the imaging input alongside much patient data including age, weight, menopausal status, detailed family history and breast density. Standalone, mammograms attained an AUC of 0.68 as an imaging-only deep learning model. Using only risk scores based on the non-imaging data attained an AUC of 0.67 through a logistic regression model. A modest improvement was found with fusion, with an AUC of 0.70. This fusion model used extracted features from the imaging deep learning model, fusing the risk factor data through concatenation. This process used both linear and rectified linear unit nonlinearities “to fuse the information” together.

Spasov et al. [73] used both MRI and various non-imaging data including patient and generic data to predict Alzheimer’s disease. A CNN model was implemented in this work where learned features were concatenated within the network and further trained throughout the architecture. Various networks were used here both to extract features from imaging and non-imaging data and to merge different features throughout. This network achieved near-perfect accuracy.

Late fusion takes the probabilistic predictions from multiple, individual models to curate the final decision. This is also known as decision-level fusion. Separate models are trained with the different modalities and a final decision is an aggregation of the predictions including averaging or majority voting.

Reda et al. [74] performed a small scale (18 patients) late fusion study on prostate cancer data, using features extracted from diffusion-weighted MRI imagine across various b-values (from Non-Negative Matrix Factorisation) as the medical imaging input and results from PSA blood tests which are a routine indicator conducted by clinicians. Single modality results were used as inputs for a larger classifier. Standalone outcome measures in this work were 77.78% accuracy for PSA screening results (k-Nearest Neighbour) and 88.89% accuracy for MRI (using Random Forest). Using data fusion with both imaging features and PSA blood test results, the accuracy increased to over 94% using a stacked auto-encoder.

Rogova and Stomper [75] use an “intelligent voter” system across two separate neural network classifiers – one with human-provided information, the other intensity-based features from mammograms, to assess calcifications. This work curated software which considered the subjectivity of human decision making and made allowances for this where different people may make a different judgement. Single modality classifiers held high sensitivity but low specificity, while the hybrid system retained the high sensitivity and improved the specificity of both radiologists and the single modality system.

Lederman et al. [76] assessed various classifiers to discriminate between high or low-risk breast cancer patients using spectroscopy. Patients were grouped based on their BIRADS score, with higher scores leading to higher risk grouping. In this work, three classifiers were tested independently – a neural network, a support vector machine and a gaussian mixture model. Although the best classifier was the neural network with an AUC of 0.81, each classifier selected three different feature subsets leading the authors to consider fusion methods. The best performing fusion methods attained AUCs of 0.84 through the combination of probabilistic output scores from all three standalone methods. The discussion noted the weighted sum rule – a linear combination of the scores of the three classifiers – was the most consistent and robust performer.

Sehgal et al. [77] used decision-level fusion across a neural network and support vector machine classifier using DNA microarray data for ovarian or breast cancer classification. This work proposed Decision Based Fusion using Stacking (DBFS) to perform the fusion methodology, using cross-validation to calculate posterior probabilities after the original classification process. This work improved the classification accuracy dramatically compared to the standalone classification process.

Majner et al. [78] used a late fusion process to classify skin lesions. Two support vector machine classifiers were used – one for deep learning features and another for hand-crafted features. Both classifiers operated separately, with the probabilistic output of both compared and the maximum score leading to the final

combination. The standalone classifiers attained accuracies of 0.794 and 0.805, while the combined classifier attained an improved score of 0.826.

Yoo et al. [79] used both joint and late fusion to predict the status conversion to multiple sclerosis within two years. MRI imaging and patient data were fused using CNN extracted features. The patient data included information of the patient's disability status and onset, and location of clinically isolated syndrome event for context. The late fusion model averaged the probabilities from the standalone models, attaining an AUC of 0.724. The joint fusion model utilised a distance transform on the MRI images, pre-training of the CNN and all patient-level data and attained the highest predictive capability of all tested models with an AUC of 0.746.

Huang et al. [80] compared all three types of multimodality data fusion listed here. To detect the presence of pulmonary embolism, this study made use of CT scans and patient electronic health records, assessing various methodologies and data combinations within early, joint, and late fusion. This includes concatenating electronic health record data or classifying them separately before combining them with deep learning features for the classification process. Convolutional features from the PENet deep learning classifier were attained, to satisfy the use of deep learning in this work.

The best model in this work, attaining an AUC of 0.947 was the "Late Elastic Average" fusion model, which averaged probabilistic predictions from both the electronic health records using a deep neural network, ElasticNet, and extracted features from the PENet for the CT scans. An early fusion model, combining extracted features and then classifying attained an AUC of 0.899 while a joint fusion model performed slightly worse at 0.893. This study highlighted the need for utilising both imaging and important clinical data which is studied later in this thesis.

The presented work on deep multimodality fusion in this thesis looks to combine deep learning features, hand-crafted statistical texture features and lesion-level metadata with the aim of improving the predictive capability of a classifier. This

forms part of the package of proposed work of machine learning-based tools for cancer triage and classification.

2.4 Summary

The literature review has identified areas which this thesis attempts to contribute to.

The Fisher Information Network methodology at present reviews dataset structures. It is intended that this will be developed in this thesis to build on work conducted in a cancer diagnosis application. This does not take new/unseen test cases into account, only training cases to represent a dataset structure. Further, only tabular data is used. This thesis will attempt to represent images rather than tabular data as well as projecting test cases within a trained embedding.

Work on interpretability of black-box machine learning algorithms is wide-reaching with applications that are model-agnostic and specific to certain architectures. This thesis will attempt to contextualise the decision-making process of a classifier, which as identified in the literature review can sometimes be limited with other approaches. Under a clinical setting this could go hand-in-hand with clinical impact.

Multimodality data fusion is an area of growing popularity in the literature, particularly using deep learning within the framework. The literature review has identified that much of the work available produces impact within the machine learning methodology over clinical contributions. This thesis will look to use a machine learning workflow to replicate the thinking of a clinician who will use multiple data sources to support a decision.

3 Dataset

This chapter describes the publicly available CBIS-DDSM dataset which has been used throughout this thesis. The Curated Breast Imaging Subset of DDSM (CBIS-DDSM) [25] is a standardised version of the Digital Database for Screening Mammography (DDSM) [81], [82]. The DDSM is a rich mammography database, containing 2,620 scanned film mammography studies with verified pathology information. The CBIS-DDSM includes a subset of the DDSM data selected and curated by a trained mammographer, for which a description of the process is described in this chapter.

Within the dataset, updated regions of interest (ROI) from the images have been provided and the pathologic diagnosis is also included. This dataset contains both breast masses and calcifications and holds a pathology label for each case, which is either “benign” or “malignant” with verified pathology information.

3.1 From DDSM to CBIS-DDSM

Within mammography there are few standard and well-known evaluation datasets. Until more recently, many CAD applications in the literature evaluate their methods on a private dataset. Available public datasets include the DDSM and the Mammographic Imaging Analysis Society database (MIAS) [83] however these hold limitations such as their size and accessibility. The DDSM dataset holds 2620 scanned film mammography studies across “normal”, “benign” and “malignant” labels verified pathologically [81], [82].

As the DDSM dataset holds many scans that were attained at hospitals with ethical approval alongside ground truth labels, the curators of the CBIS-DDSM dataset realised the value in updating and modernising the dataset. For example, the DDSM files were in older-style files that are out of date. Furthermore, annotations for the region of interest were for general position of lesions rather than exact locations. The CBIS-DDSM dataset holds the files in the more traditional DICOM file format [84] as well as improved segmentation for more accurate ROI details, using a lesion segmentation algorithm [85], verified by a trained mammographer.

One noteworthy loss of the move from DDSM to CBIS-DDSM, is that age was not transferred. In these application areas, age can be a useful data aspect to hold due to the relation between survival and age for breast cancer diagnosis [86].

3.2 Types of lesions – masses and calcifications

A breast mass is a localised lump or swelling in the breast and tends to be denoted by its size, location, shape, and margins. Depending on its criteria across the shape and margins the likelihood of malignancy can be defined and descriptions of these and their link to malignancy likelihood follow. Figure 3-1 shows an example of a breast mass.

Breast calcifications are calcium deposits, presented as “milk spots” that appear in mammograms as small white spots. They tend to appear clustered and are analysed according to the distribution, size, and shape of their appearance. These are mainly benign and usually signpost to a manifesting breast mass. A description of the calcification types and distribution follows, including the categories of these that are more likely to be benign or malignant. Figure 3-1 shows an example of breast calcification.

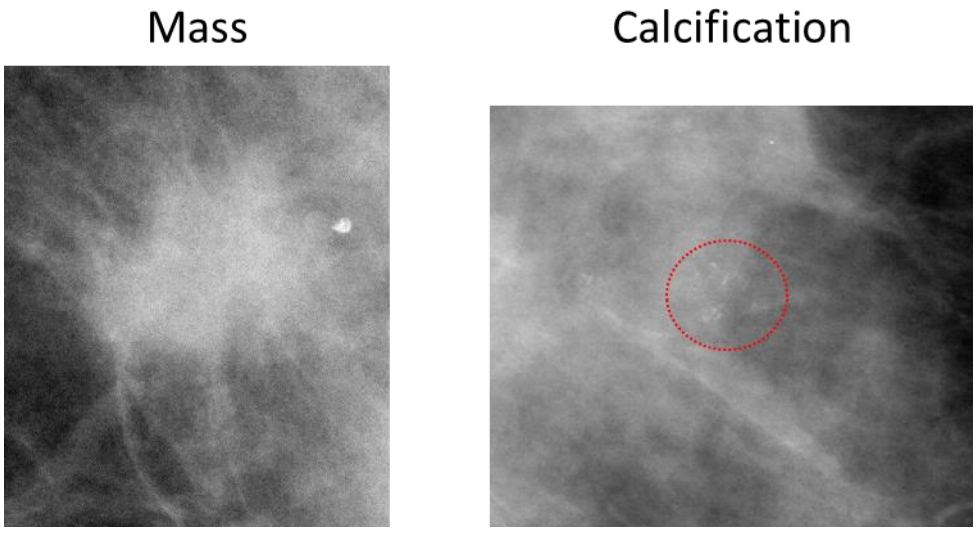


Figure 3-1: Example of a mass and a calcification, both in region of interest images. The calcification image is annotated to show the calcification (milk spots). Images from CBIS-DDSM dataset [25].

3.3 Data available within CBIS-DDSM

3.3.1 Full mammogram images

Each patient in the dataset holds mammograms across two views. These are through both the craniocaudal (CC) and mediolateral oblique (MLO) views of the breast. These views are standard projections in mammography screening. It is worth noting that there may be more than one lesion for a given patient and further within the same breast. It is known that calcifications can be an initial sign of the development of a mass [87]. Figure 3-2 shows a representative case from the CBIS-DDSM dataset of full mammogram images from the same patient, both the CC and MLO views.

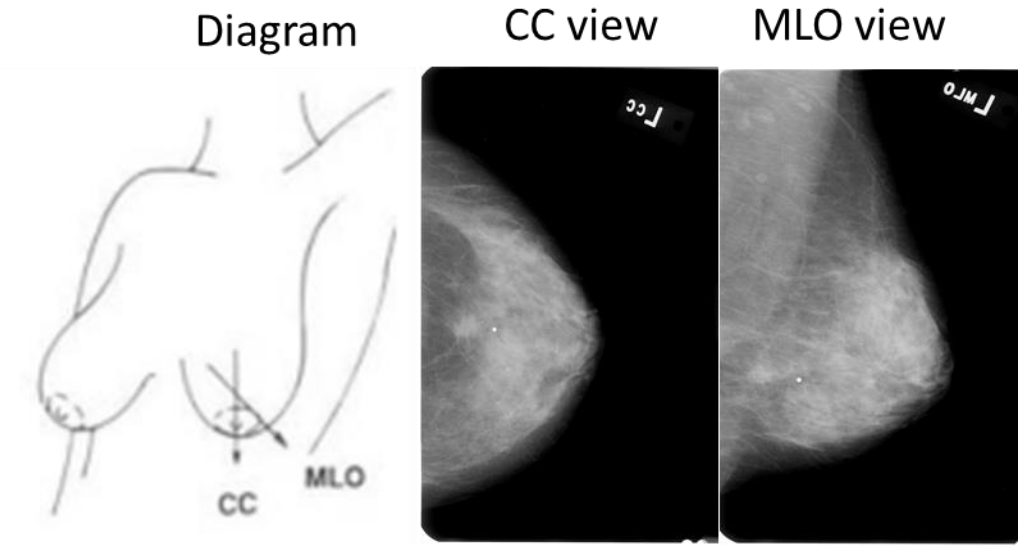


Figure 3-2: The two standard mammography views available in the CBIS-DDSM dataset are the CC and MLO views. The "CC view" and "MLO view" images here are for the same patient in the CBIS-DDSM dataset. Data images from CBIS-DDSM dataset [25].

3.3.2 Region of interest patch images

For each given lesion, an ROI patch is provided with the associated mammogram. This is a crop of the image to show the mass or calcification more directly. An example of what is provided with a given lesion case is provided in Figure 3-4.

3.3.3 Associated metadata

3.3.3.1 Metadata for all lesions

Much information about abnormalities present in a mammogram can be summed up by a radiologist, usually a pre-pathology sample, through the Breast Imaging Reporting and Data System (BI-RADS) score [88]. This score was originally developed for mammograms although its use has since been extended to MRI and breast ultrasound. It allows for images to be grouped into well-defined categories. As it is pre-pathology, one may see it as a "pre-pathology triage" system. It is of more benefit to radiologists than patients.

The scores that can be assigned are between 0-6, where BI-RADS 0 is an incomplete assessment and more mammograms are required, and BI-RADS 1 and 2 are negative and benign respectively with a near-0% likelihood of malignancy. BI-RADS 3 is "probably benign", where the likelihood of malignancy begins to appear but remains low ($\leq 2\%$) and would require short-interval follow-up. BI-RADS 4 denotes a suspicious lesion in the imaging and these cases would be pushed forward for a tissue diagnosis alongside BI-RADS 5 lesions where the likelihood of malignancy is over 95% – further categorisation can occur here, it does not in the CBIS-DDSM dataset. The BI-RADS 6 score denotes a known, biopsy-proven malignancy requiring surgical excision when appropriate (this score does not appear in the dataset but is included for completeness).

A score for breast density is provided in this dataset, as a score between 1-4, with 1 denoting lower density. Previously, breast density has been the subject of machine learning-based classification in the literature [18]. It can provide a useful indicator of how difficult it is to locate a breast lesion [89], [90], although it is generally ill-advised against for making decisions [88].

A subtlety score between 1 and 5 has been assigned to each lesion in the CBIS-DDSM dataset. Little description about this score is provided and so has not been used within models, however it is possible to use it after the analysis to reveal further insights.

In the application areas studied throughout the thesis, where metadata has been used it has been applied differently. This is to show a range of applications for this data subset and to exploit the information provided by the attained results. On a clinical view, this could provide further insights into the findings, strengthening the proposed methodologies. For each concept it will be clearly defined what data has been used and whether it was for the classification, or post-processing analysis.

3.3.3.2 Calcifications

Calcification type. Large rodlike, coarse, skin, round and regular, eggshell and milk of calcium calcification types are typically benign. Amorphous calcifications are of concern. Pleomorphic, and fine linear branching calcifications are of higher probability of malignancy.

Calcification distribution. The names of the categories of calcification distribution lend well to their visual descriptor, with examples shown in Figure 3-3. Clustered calcifications refer to groups of five or more within a small tissue volume and can be either benign or malignant. Segmental calcifications suggest malignancy and are distributed in a duct and its branches. Regional distributions occupy a larger volume of breast tissue and are either malignant or benign. Diffusely scattered calcifications are distributed randomly throughout the breast and are almost always benign. Linear calcifications are suggestive of malignancy.

Table 3-1 lists the calcification types and distribution categories available in the CBIS-DDSM dataset.

3.3.3.3 Masses

Mass shape. There are eight categories of mass shape present in this subset of the CBIS-DDSM dataset. In general, they tend to describe the shape of the lesion with drawn examples of these shown in Figure 3-3. Architectural distortions are defined by the Breast Imaging Reporting and Data System (BI-RADS) system as an appearance in which “the normal architecture of the breast is distorted with no definite mass visible.” Asymmetries detect differences between the breasts, as the internal structure of the two breasts for a given person are very similar.

Mass margins. There are five categories of mass margins, which defined the border of the mass and is an important aspect of determining the classification of a lesion [91]–[93]. Circumscribed margins are well-defined and are sharply demarcated with a notable change from normal tissue to the lesion. The likelihood of malignancy tends to be lower. Ill-defined margins are not well defined and can be scattered. Obscured margins are hidden from view by normal tissue. Spiculated margins are marked by radiating thin lines. These tend to have a higher likelihood of malignancy. Micro-lobulated margins have small, undulating circles along the edges of the mass.

Table 3-2 lists the mass shapes and margin categories available in the CBIS-DDSM dataset.

Calcification type	**Calcification distribution**
Amorphous	Clustered
Coarse	Linear
Eggshell	Segmental
Dystrophic	Diffusely scattered
Pleomorphic	Regional
Punctate	
Milk of calcium	
Fine linear branching	
Large rodlike	
Lucent centre	
Round and regular	
Skin	

Table 3-1: Categories of the calcification metadata within the CBIS-DDSM dataset

Masses:

Mass shape	**Mass margins**
Architectural distortion	Circumscribed
Asymmetric breast tissue	Ill defined
Focal asymmetric density	Obscured

Irregular	Spiculated
Lobulated	Microlobulated
Lymph node	
Oval	
Round	

Table 3-2: Categories of the mass metadata within the CBIS-DDSM dataset

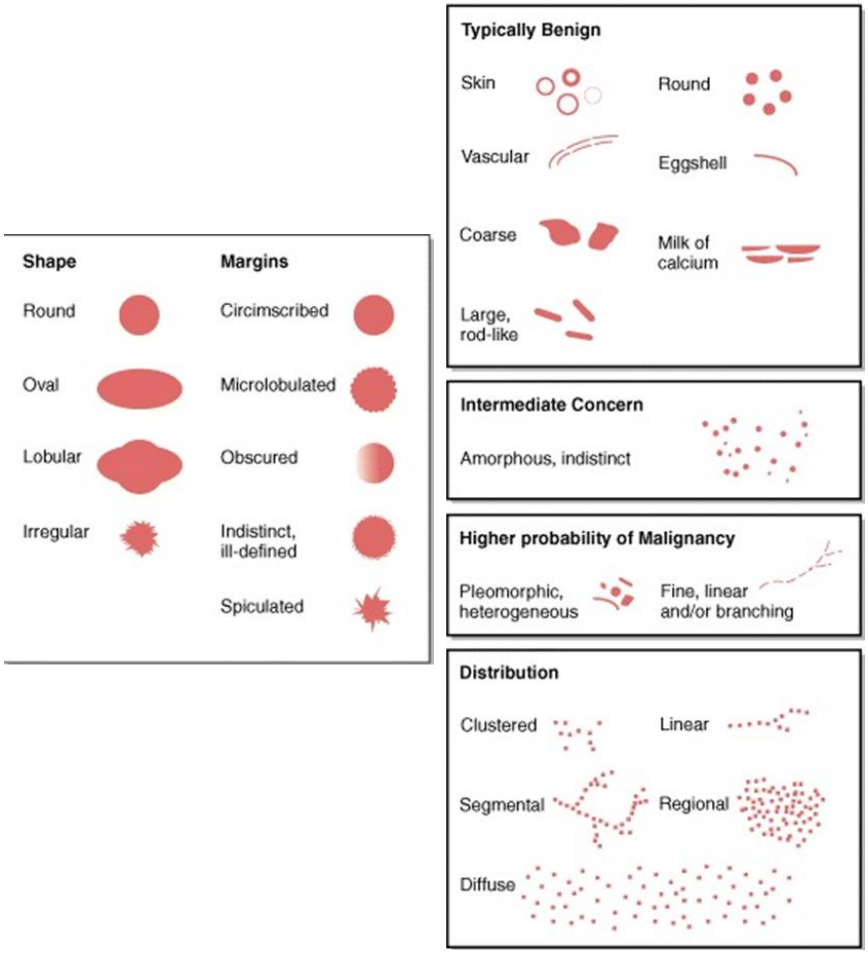


Figure 3-3: Terminology and visual representation of masses (left) and calcifications (right). Taken from The Abnormal Mammogram [92]

A representative example of the available data - full mammogram images, cropped ROIs, patient information and associated metadata – is shown in Figure 3-4.

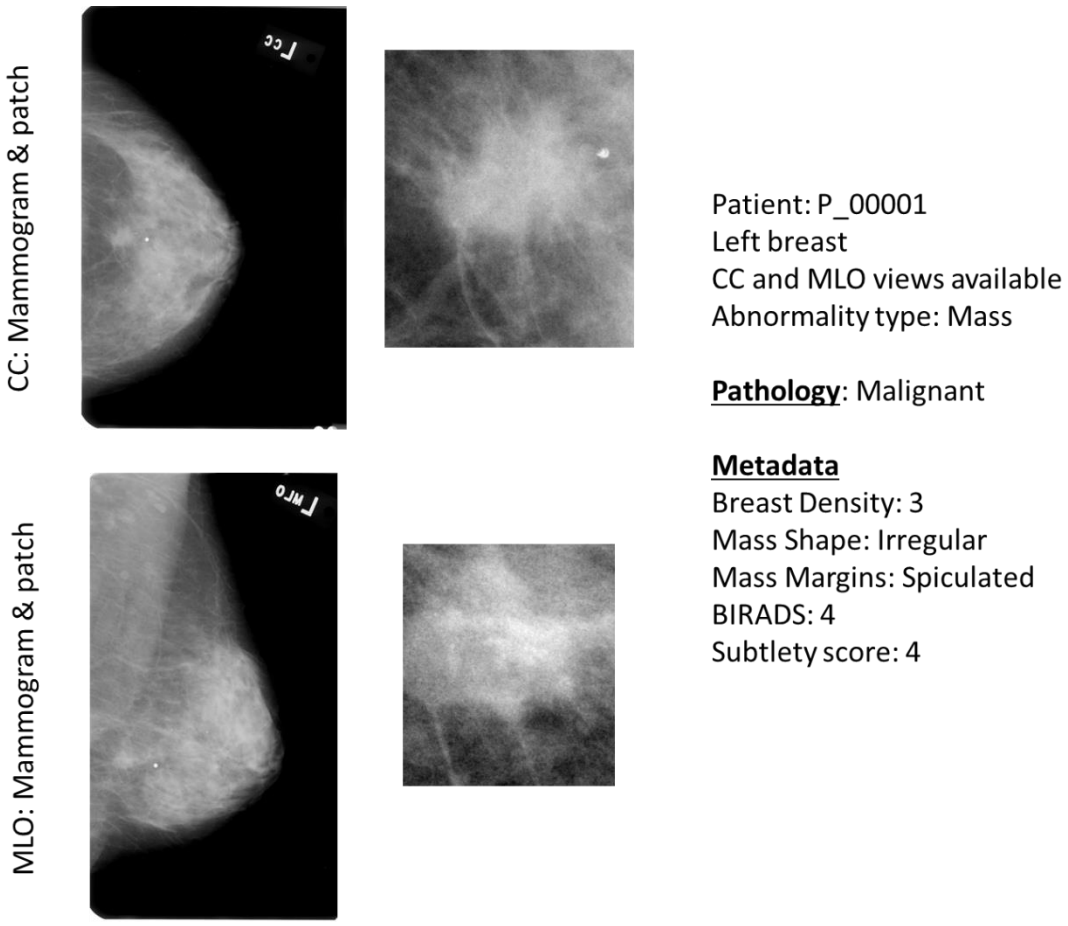


Figure 3-4: Example of a given lesion - the mammogram patches taken and the ROI alongside the associated metadata.

3.4 Data derived from images in the dataset

Parts of the content in the thesis look at forms of "multimodality" analysis.

3.4.1 Statistical texture features

A total of 130 statistical texture features – or "hand-crafted features" - from each CBIS-DDSM ROI patch were extracted and have been used in mammogram classification in the literature [94]. These ROIs were enhanced using Contrast Limited Adaptive Histogram Equalization [95] to reduce noise within the images. This method divides the ROI into equally sized contextual regions, applies histogram equalisation on each region, limits the histogram by a clip level and redistributes the clipped amount throughout the histogram. The pixel value is then obtained by histogram integration. Radiomics [96] looks at the extraction of numerous quantitative imaging features. Features described here are extracted from defined regions of medical images to "provide accurate risk stratification by incorporating

the imaging traits into predictive models for treatment outcome and to evaluate their added value to commonly used predictors." [96] Their use in this thesis is to support the evidence in the literature of how they apply within decision support systems.

Textural features quantify information about spatial distribution of tonal variations within a band [97]. Statistical methods analyse the spatial distribution of greyscale values, by computing local features at each point in the image and deriving a set of statistics from the distributions of local features [98]. Depending on how many pixels define the local feature, statistical features can be first order (one pixel), second-order (two pixels) or higher-order (three or more pixels). As first-order statistics ignore spatial interaction between image pixels, it is proposed to use both first and second-order statistics to represent the images. The second-order features extracted are Grey Level Co-occurrence Matrix (GLCM) and Grey Level Run Length Matrix (GLRLM), at angles 0, 45, 90 and 135. [97], [99]

3.4.2 CNN features

Some application areas within the thesis refer to the use of CNN features as a dataset. This refers to a convolutional neural network model curated by Shen et al. [100], a Resnet50-based model. This has been used to study and exploit the strengths of a well-performing deep learning architecture. This model is a strong performing model in the literature with the CBIS-DDSM dataset as well as publicly available. This thesis studies the patch classifier, using mammogram patches provided by the author using a sliding window across the mammograms with information available about the location of the given lesion. The output of the penultimate layer of the model has been used as a feature extractor of the network and these outputs are used as the "CNN features" throughout the thesis.

4 Fisher Information Network methodology

This chapter describes the Fisher Information Network methodology. This is done to support the following chapter in the thesis which studies two different applications of the FIN approach including a "patient-like-me" method to study new patients. The methods in this chapter build on work proposed by [30], [101].

Figure 4-1 shows a diagram of the process from data pre-processing to the application stage, for which the main Fisher Information Network methodology will be described here. After data pre-processing and development of a multilayer perceptron (MLP) classifier, the probabilities derived from the classifier are used to derive the Fisher Information metric, which leads to the estimation of the pairwise distances within the Fisher manifold. Then, using multidimensional scaling, each given case can be visualised together through a form of dimensionality reduction. This leads to projecting the new test cases, unseen to the original classifier, leading to an application-based analysis in the following chapter.

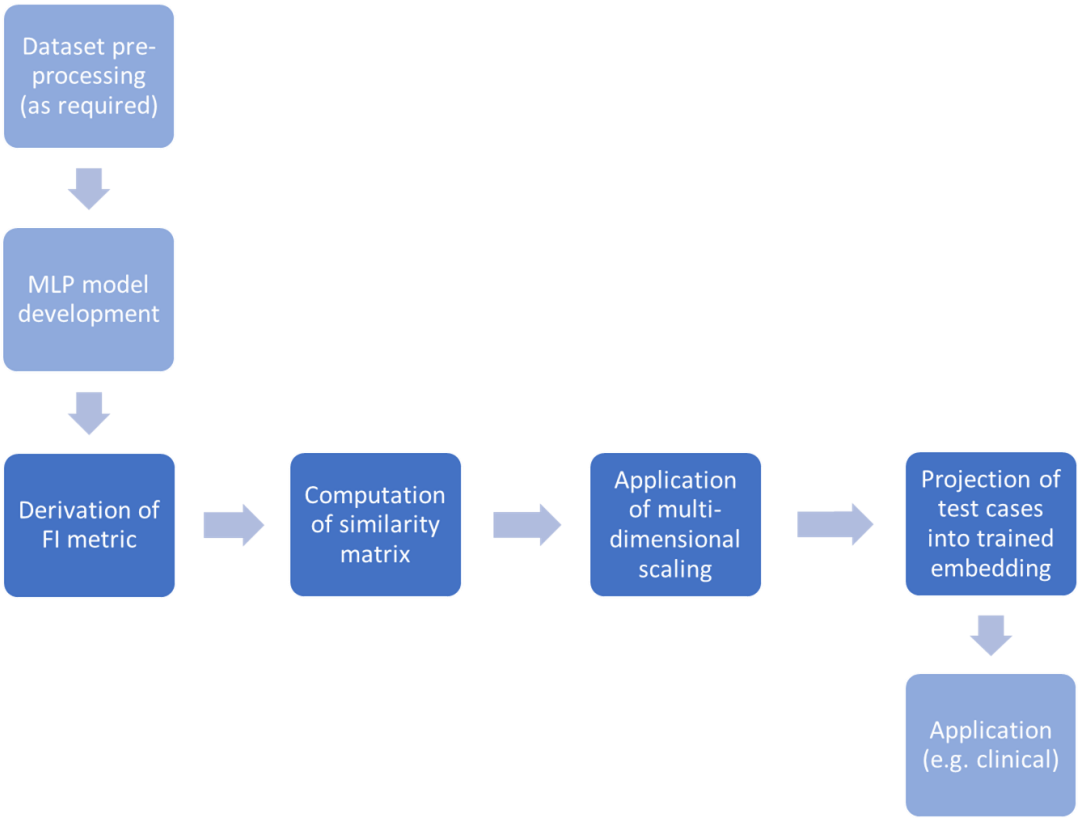


Figure 4-1: Diagram of the Fisher Information Network methodology. The four centred boxes detail the main stages of the FIN approach.

4.1 Multilayer Perceptron for predictive probabilities

For this work, the MLP has been used as the discriminative model. A type of neural network built from neurons [102], the power of the classification is built from the links between the several in a given MLP network. The 'learning' occurs through the units adapting their weights of the connections between themselves as the network is trained with new training data. Figure 4-2 shows a representation of an MLP, for a binary classifier. Input data is entered at the input layer, pushed to the hidden layer.

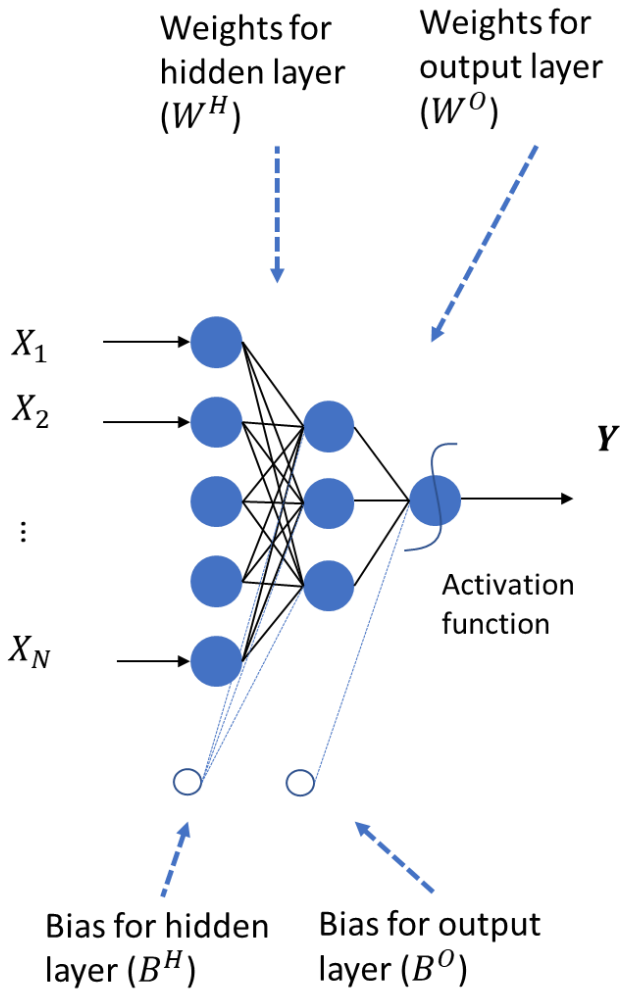


Figure 4-2: Representation of an MLP, with weights and biases defined. Input data entered at the start, pushed to the hidden layer. Iteratively, as more training data enters the network, the weights and biases adapt to generalise to the data. With the correct setup (number of units, activation)

Iteratively as more training data is inputted, the weights and biases adapt to learn how to discriminate between two classes. The hidden layer can extract higher-order statistics from the input, generating non-linear intermediate signals. Outputs from

the hidden layer are combined in the output layer, leading to the outputs of the MLP:

$$a(\boldsymbol{x}) = \boldsymbol{W}^{O}\boldsymbol{\theta}(\boldsymbol{W}^{H}\boldsymbol{x} + \boldsymbol{B}^{H}) + \boldsymbol{B}^{O}$$

Equation 4-1

Where x is the input data, $\boldsymbol{W}$and $\boldsymbol{B}$ are the weights and biases of the hidden (H) layer and the output (O) layer, respectively. θ is the sigmoid function, used as the probability estimator for a binary classifier:

$$\theta(z) = \frac{1}{(1 + e^{-z})}$$

Equation 4-2

For multi-class classification where J reflects the number of possible classes, θ in Equation 4-1 can be replaced with the SoftMax activation function:

$$p(c_j|x) = \frac{e^{a_j(\boldsymbol{x})}}{\sum_{k=1}^{J} e^{a_j(x)}}$$

Equation 4-3

4.2 Derivation of the FI metric

The probability densities of the classes estimated with an MLP can then be used to calculate pairwise distances, producing the Fisher distance matrix. In this step, the FI metric is derived from the Fisher distance matrix. It is obtained by differentiating the logarithm of the conditional probability $p(x|\theta)$ with respect to x and summing over all possible classifications. The metric defines a Riemannian space where the distances are a measure of similarity between the respective probability distributions, as shown in Equation 4-4.

$$FI(x) = E_{p(x)}\{(\nabla_x log\ p(x))^T(\nabla_x log\ p(x))\} = -E_{p(x)}\{\nabla_x^2 log\ p(x)\}$$

Equation 4-4

Where $E_{p(x)}$ denotes the expectation over the density function $p(x)$ and ∇_x is the gradient with respect to x.

The distance between two neighbouring points is then calculated by the quadratic differential form, as shown in Equation 4-5, known as the “straight path” approach:

$$d(x, x + \Delta x)^2 = \Delta x^T FI(x_A)\Delta x$$

Equation 4-5

Here, the "straight patch" approach considers the shortest path (distance) to be a straight line between the two points, which provides an approximation of the integral along that path. This assumes that the FI matrix, $FI(x)$, is constant throughout – where $FI\big(x(t)\big) = FI(x_A)$.

The integral can be approximated using interpolation methods. For this, the FI matrix is evaluated at T points along the defined straight path. This can approximate the true length of the straight line in this approach, through the choice of T. As recommended by [30], T = 10 is suggested as per empirical experiments.

$$d_T(x_A, x_B) = \sum_{t=1}^{T} d_1(x_A + \frac{t-1}{T}(x_B - x_A), x_A + \frac{t}{T}(x_B - x_A)$$

Equation 4-6

The shortest path is estimated using Floyd-Warshall algorithm [103], [104].

4.3 Computation of the similarity matrix

The similarity matrix is then computed from the distance matrix generated previously, using a Gaussian radial kernel, resulting in an adjacency matrix that defines the network structure (see Equation 4-7). Here, distances are transformed into similarities

$$A_{ij} = e^{-\frac{\Delta x^2}{\sigma_G^2}}$$

Equation 4-7

Where Δx^2reflects the distance between two points, x_i and x_j within the Fisher metric, σ_G controls the influence of locality in the generation of the weights of network connections and is determined using a heuristic method. It is taken as the average pairwise distance between points belonging to same predicted label. A small value of σ_G results in closer points having significant connection weights, while larger values reduce this effect and produce meaningful values for points that are further away.

The distances $d(x_i, x_j)$ lead to the similarities, A_{ij} w.r.t the estimate of the predictive probabilities. The adjacency matrix, A_{ij}, contains the structure of the Fisher network. Each point holds a similarity to every other point in the network, which leads to the aforementioned "distances" between points. Through the use of MDS, it will be possible to project these similarities into a visualisation by condensing this high dimensionality matrix into fewer (usually 2 or 3) dimensions.

4.4 Application of multidimensional scaling

The use of multidimensional scaling is to embed the Fisher manifold into a lower-dimensional space, allowing for visualisation of the Fisher manifold's structure. This transforms the pairwise distances held in the Riemannian manifold into coordinates which are embedded in a Euclidean space.

Multidimensional scaling [31] searches for a lower-dimensional space which is usually Euclidean, in which each of the points within the space represents a given data point. Each point represents a given case in the input data. The overall aim is that the distances between each point in the feature space, reflect the (dis)similarities.

After the computation of the similarity matrix, MDS produces a representation of the patients in a low-dimensional Euclidean space, such that the distance between two cases in the lower dimension Euclidean space approximates as closely as possible the distance between the respective cases in higher dimension Riemannian space.

MDS [31] uses a matrix of pairwise dissimilarities between cases to produce a reflection of the instances to show the distances approximate as closely as possible the dissimilarities between the corresponding cases in the original matrix. For visualisation purposes, it aims to find a configuration of n points in a (usually Euclidian) space so that each object is represented by a point in the space.

4.4.1 Minimising the stress loss function

To find a mapping, ϕ, of the dissimilarities, $\delta_{rs,i}$, giving rise to the set of disparities, $\widehat{d_{rs,i}}$, the task is defined as:

$$\phi[\delta_{rs,i}] = d_{rs,i}^2$$

Equation 4-8

Where $\hat{d}_{rs,i}^2$ are least squares estimates of $d_{rs,i}^2$ which are obtained through minimising the loss function:

$$SS = \sum_r \sum_s \sum_i \left(d_{rs,i}^2 - \hat{d}_{rs,i}^2\right)^2$$

Equation 4-9

In this work, the squared stress metric (SS) is used as a measure of the goodness of fit of the approximation of the original dissimilarities. It is minimised using an alternating least squares algorithm. Known as ALSCAL (Alternating Least squares SCALing) [105], where we consider the SS metric as a function of the similarity coordinates matrix, $\boldsymbol{X}$, a matrix of weights, $\boldsymbol{W}$ and the disparities, $\widehat{D}$: $SS(\boldsymbol{X}, \boldsymbol{W}, \widehat{D})$, the algorithm is as follows:

1. Find an initial configuration of both $\boldsymbol{X}$ and $\boldsymbol{W}$**.**
2. Perform optimal scaling to calculate D, $D*$ and normalise.
3. If the SS metric has converged, terminate.
4. Perform model estimation: minimise $SS(\boldsymbol{W}|\boldsymbol{X}, D^*)$ over W; then, minimise $SS(\boldsymbol{X}|\boldsymbol{W}, D^*)$ over **X**.
5. Go to step 2.

The optimal scaling phase takes the calculated distances, $d_{rs,i}$ from the current coordinate and weight matrices, $\boldsymbol{X}$ and $\boldsymbol{W}$. Disparities, $\hat{d}_{rs,i}^2$, are then calculated. Placing all disparities in a vector, $\hat{d}$ and the distances into d, then

$$\widehat{\boldsymbol{d}} = \boldsymbol{E}\boldsymbol{d},$$

Equation 4-10

Where $\boldsymbol{E} = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T$, with **Z** depending on the type of transformation. **Z** is a matrix of variables indicating which distances must be tied to satisfy the measurement conditions.

SS can now be denoted as

$$SS = \boldsymbol{d}^T(\boldsymbol{I} - \boldsymbol{E})\boldsymbol{d}$$

Equation 4-11

The model estimation phase looks for the least squares estimated of the weight matrix, **W**, for the current disparity values, $\hat{d}_{rs,i}$an coordinates **X** of the points in the group space. The least squares estimates are then found of **X** for the current disparity values and weights **W**.

For the first minimisation, let the $\frac{1}{2}n(n-1)$ quantiles $(x_{rt} - x_{st})^2$ make up the t^{th} column of a matrix, **Y**. A similar $\frac{1}{2}n(n-1)$ p matrix, $\boldsymbol{D}^*$ is composed of the disparities $\delta^2_{rs,i}$. Then, SS can be written as:

$$SS = (\boldsymbol{D}^* - \boldsymbol{W}\boldsymbol{Y}^T)^T(\boldsymbol{D} - \boldsymbol{W}\boldsymbol{Y}^T)$$

Equation 4-12

and hence,

$$\boldsymbol{W} = \boldsymbol{D}^*\boldsymbol{Y}(\boldsymbol{Y}^T\boldsymbol{Y})^{-1}$$

Equation 4-13

In the following chapter, visualisations from the MDS process facilitates the analysis of individual cases using a 'patient-like-me' approach, leading to clinical impact. Where a case exhibits similar characteristics to another case, they will appear close together on the visualisation. Throughout this work, MDS looks to reduce high-dimensional data into a 2- or 3-dimensional representation, while still showing the approximation of dissimilarity through distances between points. The intention of this is for unseen test cases to be embedded into the manifold, leading to new cases that share similar characteristics appearing in the same area as a trained case from the trained classifier. A description of this method follows.

4.4.2 Projection of test cases into trained embedding

The distances of each test point from each training point are also computed. However, for these distances to be appropriate to project the test cases into the trained embedding, a procedure to compute the weighted average of the training coordinates is required, to ensure the weights are proportionally inverse to the distance.

For this to be calculated, a length scale parameter is calculated, σ_G. This step transforms the manifold distances into a similarity network using a Gaussian kernel based on Equation 4-7. The similarity network is defined through A_{ij} which is symmetrical. The method used to calculate σ_G is a heuristic approach using the average intra-label distances in the Fisher manifold [101].

The intra-label distances represent the average pairwise distances of the manifold which belong to the same label.

$$dist_{intraLab}(x_i, x_j) = \frac{\sum_{i=1}^{N} \sum_{j \ni (j>i) \wedge (\mathrm{L_i}=L_j)} dist(x_i, x_j)}{\sum_{i=1}^{N} \sum_{j \ni (j>i) \wedge (\mathrm{L_i}=L_j)} 1}$$

Equation 4-14

By using the Gaussian kernel as the similarity measure, where values of $A_{i,j}$ are between 0 and 1, a constraint may be imposed that the intra-label distance must be equal to a relatively high value of similarity, setting $A_{th} > 0.5$.

$$A_{(i,j) \in \mathrm{intraLab}} = e^{\left(-\frac{r.dist_{intraLab}(x_i, x_j)}{\sigma_G}\right)^2} \leftarrow A_{th}$$

Equation 4-15

Through imposing this condition, σ_G is determined within a small range of values (the r value is a multiplying factor for network granularity).

$$\sigma_G \leftarrow \frac{r.dist_{intraLab}(x_i, x_j)}{\sqrt{\ln(A_{th})}}$$

Equation 4-16

This sets σ_G, allowing for the adjacency matrix A_{ij} to be calculated for the test cases.

$$A_{ij}(test) = \exp \frac{-\left(\frac{dist(x_i, x_j)}{\sigma}\right)^2}{\sum \left(\frac{dist(x_i, x_j)}{\sigma}\right)^2}$$

Equation 4-17

The transpose of these values is then multiplied by the scaled coordinates of the training data. This brings the test data coordinates into the same scale as the training coordinates. At this stage, it is possible to project the test cases, using their coordinates, into the trained embedding as they are now within the same scales as each other. This leads to more practical applications, for which two examples of clinical data on breast cancer are detailed in the following chapter.

4.5 Chapter summary

This chapter has provided a description of the Fisher Information Network methodology, which will be applied to real-world data in the following chapter. This involves thedevelopment of an MLP, derivation of the FI metric, computing the similarity matrix, leading to multidimensional scaling and projection of test cases into the manifold. Using posterior class probabilities from an MLP and its weights and biases, this method looks to take what has been inherently learned by the classifier and visualise the data allowing for a new understanding of the algorithm's output.

The manifold reflects the information learned by the classification model on which it is based. Once the Fisher manifold is in the Euclidean space, MDS can be used to visualise the structure of the embedding, leading to more practical applications. Visualisations can assist both experts and the vernacular in understanding the workings of a machine learning classifier. Further, the projection of new test cases is a novel research point in this thesis. This can exploit the predictive capability of a "black box" machine learning method such as an MLP. In the following chapter, these points are demonstrated.

5 Fisher Information Network – Application of a clinical problem

This chapter studies two different applications of the Fisher Information methodology on the CBIS-DDSM breast cancer dataset. Both look at a "patient like me" approach for new test cases. The first focusses on creating a classifier which discriminates between benign and malignant breast lesions using statistical texture features of images. The second looks at a well-performing deep learning classifier on the dataset, which is a multinomial classifier, to study the application of CNN features on the same process. This can be seen as a machine learning approach for a triage-like tool for new cases and how a trained classifier can inform clinical thinking.

5.1 Applying the FIN methodology to texture features, leading to a patient-like-me approach

5.1.1 Introduction

This section looks at curating a robust visual representation of clinical data from a neural network classifier. For this, a study of the breast cancer dataset, CBIS-DDSM, was conducted using the FIN methodology, initially using communities [106], later on improving the approach to create more reproducible data groupings. This allowed us to create more reproducible representations and applied this methodological approach to breast cancer data, specifically calcifications because it is a larger and richer dataset that holds similarities to the characteristics of the prostate cancer problems (calcifications are milk spots). The initial MLP model informs the FI metric, which takes advantage of the probabilities of class membership. The data used in this application are statistical texture features calculated from ROI image patches from mammograms.

Much of the existing work conducted in breast cancer classification looks at the discrimination between tumour and normal tissue [107]–[112] with a relatively high success rate. However, that differentiation is not the most useful since tumours can be either benign or even be at different stages of malignancy and being able to develop a better understanding of this for each individual patient is hugely

important, especially since malignant tumours would require immediate treatment. This is the reason why in this study the focus is on the development of a visualisation of the breast cancer patients' latent space using mammography images, showing a spatial representation of these patients of whom we have additional insights that can be used to better understand and diagnose new patients. In order to achieve this, in this thesis the proposed methodology uses neural networks augmented by the use of the Fisher Information (FI) metric, as a learning metric of a latent variable space [113], [114]. The FI metric is a natural statistical measure of dissimilarity for small changes between the data points (in this study, patients) according to their degree of relevance with respect to class membership. The proposed method produces a novel, low-dimensional visualisation (2D or 3D) that includes the projection of every mammographic image included in the analysis.

For the proposed methodology, a FI network is constructed using probability density estimates, which are calculated for two classes – benign and malignant tumours, namely for calcifications. As mentioned, the aim of this work is to produce a visualisation of the latent space of breast cancer patients using mammography images from where underlying patterns and structures from these patients could be grouped and understood.

It is expected that a visualisation of this latent space obtained with the FI network will help elucidate the underlying data structure, with the goal of assisting the diagnosis of new patients. The latter can be achieved by projecting new unseen instances/observations into this latent space, given that a huge amount of information can be learnt from the closest neighbours, which would be potentially relevant to those new patients – we call this a 'patient-like-me' approach. The latter has the potential to help identify not only what would be the most likely diagnosis, but also what kind of treatments or therapies were more successful for those nearby cases, which could inform and influence the decision of a clinician.

The main objective of this part of the thesis is to create a visualisation of the latent space of the cancer patients that represents the variability that can be found in the datasets, from which we gain insights on how different patients can be related,

leading to the development of a ‘patient-like-me’ approach. To achieve this, we propose a methodology that follows the steps below, also illustrated in Figure 5-1:

I) **Image enhancement**: This step involves enhancing the extracted image, using Contrast Limited Adaptive Histogram Equalisation [115], to reduce noise within the images, improving the classification process.

II) **Image representation and model optimisation**: This step involves the extraction of first and second-order features from the ROIs of the images, and then selecting the most relevant of them for the specific problem when using the MLP [116].

III) **Development of an MLP model:** The MLP is a class of feedforward artificial neural network [117], which has been utilised to estimate the probability densities of the classes. For the development of the model, the MLP parameters are set according to the optimised model obtained in the previous step.

IV) **Creation of the latent space of patients**: This step involves deriving the FI metric that amplifies distances along important directions, computing the similarity matrix, and applying a Multidimensional Scaling (MDS) [118] method to map the original data onto a Euclidean projective space.

V) **Detection of patients’ clusters**: Using k-means clustering, groupings of patients’ clusters are detected on the Euclidean projective space.

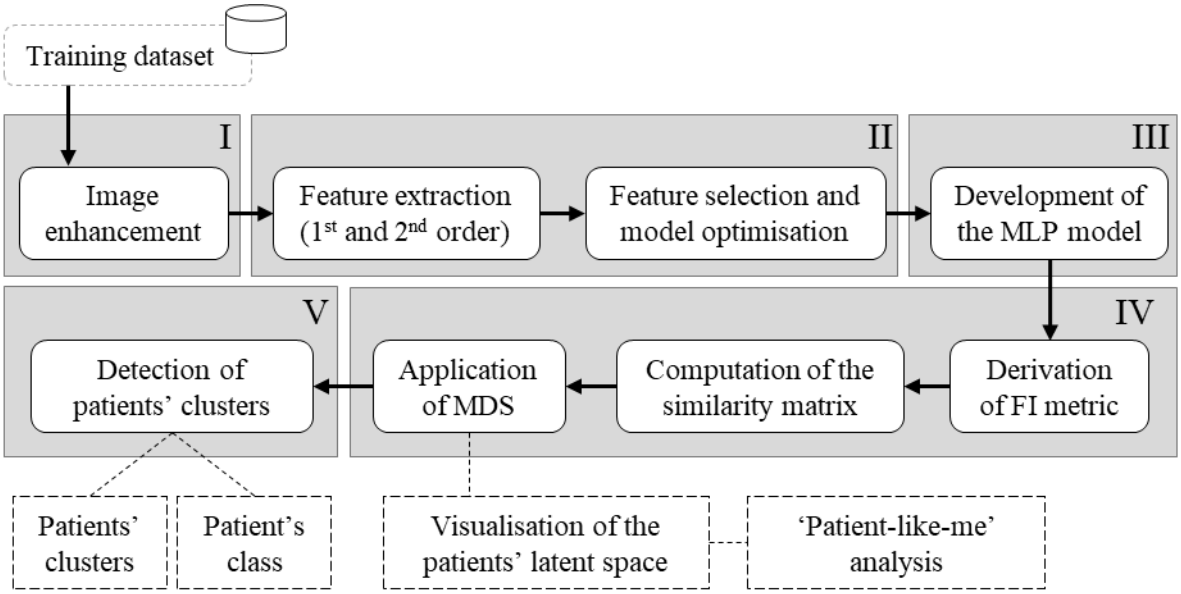


Figure 5-1: Proposed methodology: approach followed for creating the visualisation of the latent space of cancer patients to develop a 'patient-like-me' analysis

5.1.2 Methods

A detailed description of the five steps is described:

5.1.2.1 Step I: Image enhancement

Contrast Limited Adaptive Histogram Equalisation (CLAHE) [115], [119] has been used in this paper to enhance the ROIs and images to improve the contrast distribution within the images. This method divides the inputted images into contextual regions of equal size, applies histogram equalisation on each region, limits the histogram by a clip level and redistributes the clipped amount throughout the histogram. The pixel value is then obtained by histogram integration. Both CLAHE and Histogram Equalisation were explored for the study; CLAHE achieved better results and are therefore presented.

5.1.2.2 Step II: Image representation and model optimisation

Feature extraction. After ROI selection and image enhancement, we propose to extract textural features from the images. Textural features contain information about spatial distribution of tonal variations within a band [120]. They can be extracted using statistical methods, among other approaches not considered in our study such as model-based methods and transformation-based methods.

Statistical methods analyse the spatial distribution of greyscale values, by computing local features at each point in the image and deriving a set of statistics from the distributions of local features [121]. Depending on the number of pixels defining the local feature, statistical features can be first-order (one pixel), second-order (two pixels) or higher-order (three or more pixels). As first-order statistics ignore spatial interaction between image pixels, we propose to use both first- and second-order statistics to represent the image.

Feature selection and model optimisation. This stage aims at ensuring that noisy and redundant features are discarded, and only an optimal set of discriminating features are retained to represent the image. In this study, we calculated the importance of the features using Random Forest (RF) [122], since the tree-based strategies used by RF naturally ranks by how well they improve the purity of the node. Hence, depending on where the cut-off point is specified (i.e. pruning the tree below a particular node) a subset of the most relevant features can be created.

5.1.2.3 Step III: Development of MLP model

In this step, an MLP was trained to estimate the probability densities of the classes, since MLP is a semi-parametric non-linear probabilistic model of class membership, for which a FI metric can be derived [114]. The MLP was trained to classify images into the two classes (malignant and benign) using the training data. The developed MLP model was then tested on the test set, which was unseen to the training process.

5.1.2.4 Step IV: Creation of the latent space of patients

The description of the methods required for the creation of the latent space of patients in this stage – deriving the FI metric, computation of the similarity matrix and the application of MDS leading to projecting new unseen test cases – has been described in the preceding chapter in the thesis.

The obtained visualisation of the patients' latent space would facilitate the analysis of individual cases using a 'patient-like-me' approach, leading to clinical impact. Where a case exhibits similar characteristics to another case, they will appear close together on the visualisation. An aim of this study is for new, unseen cases to be

embedded into the manifold, leading to new patient cases sharing similar characteristics appearing in the same area as a patient case as per the trained model/visualisation.

5.1.2.5 Step V: Detection of patients' clusters

After creating the FI embedding with all the data points (patients) projected onto it, and mapping them onto the Euclidean space, K-means clustering was used to generate representative patients' clusters. Given that the ideal number of clusters that should be calculated is unknown, the Separation-Concordance (SeCo) framework was used [123] to find the optimal values of cluster number that would produce a clustering solution with increased reproducibility and stability, i.e. obtaining partitions more resilient to random perturbations [124]. After clusters of patients are estimated, we study their individual features to see whether we can find additional insights within each subgroup that can help us characterise the cluster. From this analysis, we can potentially identify clusters that mainly would be representing the different classes (depending on the data problem).

5.1.3 'Patient like me' approach

To satisfy the 'patient-like-me' objective of this work, test cases are projected onto the latent space of trained observations. These new test cases are expected to share similar characteristics to the cases in the training, therefore, without informing the model of their pathology, we hipothesise that these shared characteristics will be exposed. To achieve this, a similar methodology that was applied to the training data will now be applied to the test data, which is illustrated in Figure 5-2.

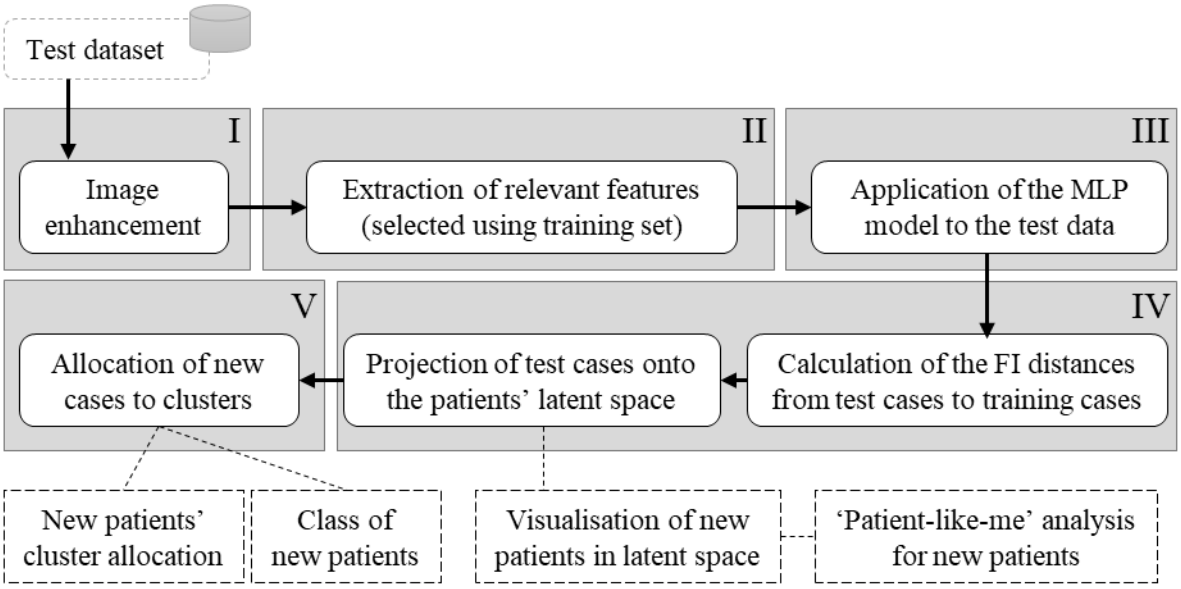


Figure 5-2: Applying the proposed methodology to a test dataset, to project new, unseen cases on an existent latent space of cancer patients

I) **Image enhancement for test cases**: The same image enhancement method applied to the training data is now applied to the test cases.

II) **Extraction of relevant features**: The relevant features selected from the nested cross-validation feature selection process are extracted from the test data to apply the test cases to the trained model.

III) **Application of MLP model to the test data:** The test data is classified against the trained MLP model.

IV) **Projecting new cases onto the latent space of patients**: The new patients' data is brought into the same projective space as the training data using each given patients' pairwise distance with respect to the training set. The test cases are then projected on the same latent space as the training cases, which will be used for reference.

V) **Allocation of new cases to clusters**: Using the previously trained clusters, the newly projected test cases are allocated a cluster. At this point, the characteristics for the test data as discovered at the training stage are assessed.

In this section, a short analysis of individual cases, to highlight how such a model can be applied to an individual patient in a care setting, will be presented and analysed.

5.1.4 Experimental settings: Implementation of the proposed method

In this study, the statistical texture features of the two views of the same breast lesion within the breast cancer dataset – CC and MLO – were concatenated.

Texture features: A list of the initial first-order and second-order features calculated can be found here [125]. The total number of features extracted was 129, as second-order features were taken at multiple angles for which the pixel to which the current pixel was compared.

Feature selection and model optimisation: In order to select an optimal set of discriminating features, we propose to use nested cross-validation (10-fold inner and 10-fold outer cross-validation) to assess the performance of an MLP model on different sets of selected features, calculated using RF. The proposed nested cross-validation would allow for a thorough and appropriate analysis of the initial MLP model, including the evaluation of different cut-off points to create a subset of features (which we limited to a maximum of 15% of the total number of variables), and a greedy search for identifying the best number of hidden nodes of the MLP (to a maximum of 10 nodes to avoid complex models and overfitting). From all these models, one of them will need to be chosen as the one that best represents the problem. In this study, in an attempt to avoid overfitting, the model to the training data, we chose as a criterion the lowest gap between the AUC of the training and the AUC of the test cases of the cross-validation. Notice that this process does not involve the independent test set, only the proportion of the training separated for test (a validation set) during the cross-validation.

MLP model: We used one hidden layer plus the output layer. The hyper-parameters of the best model identified during the model optimisation were used to develop the MLP model that was used to calculate the probability densities (i.e. probability of class membership) to generate the FI metric. This would inform the optimal number of nodes in the hidden layer of the MLP, as well as the optimal set of

features. Early stopping was implemented with a maximum number of epochs set to assist in preventing overfitting. The performance of this MLP model was tested against an independent test set.

5.1.5 Results

Firstly, the results for the breast cancer data problem, which utilises the calcification cases of the CBIS-DDSM dataset are visualised using the FIN and analysed using k-means clustering. This is followed by the evaluation of the presented results at the end of the section.

5.1.5.1 Latent space model created using breast cancer data – CBIS-DDSM

A total of 595 cases were used to train the MLP to create the latent space model, while 124 cases were used as test cases. Table 5-1 denotes the AUC of the train and test sets of the model. Furthermore, from the nested cross validation process the number of selected features and hidden nodes are listed.

Training AUC	0.828	**Number of selected features**	10 (4%)
Test AUC	0.744	**Number of hidden nodes**	5

Table 5-1: Training and testing AUC's and hyperparameter tuning results for the presented model.

Figure 5-3 shows the 2-dimensional and 3-dimensional visualisations of the calcification cases of the dataset after the FIN and the application of the multidimensional scaling process.

The SeCo framework informed the next part of the process, which was to select the most suitable number of clusters to partition the dataset into, leading to further generalisability for test data. This process denoted that 5 clusters were the most suitable number of clusters to partition the data into in this process.

Table 5-2 contains the distribution of the clusters against the class label, the pathology, for the training dataset. Figure 5-4 shows a graphical visualisation of the k-means clustering of the feature selected data in the latent space of the training dataset, with the cluster centroids marked. From this, it can be noted that Cluster 1 consists mainly of malignant cases and one benign case. Clusters 3, 4 and 5 consist mainly of benign cases, compared to a relatively small amount of malignant cases.

Cluster 2 contains a mix of both malignant and benign cases and therefore a categorisation cannot be reached in this case.

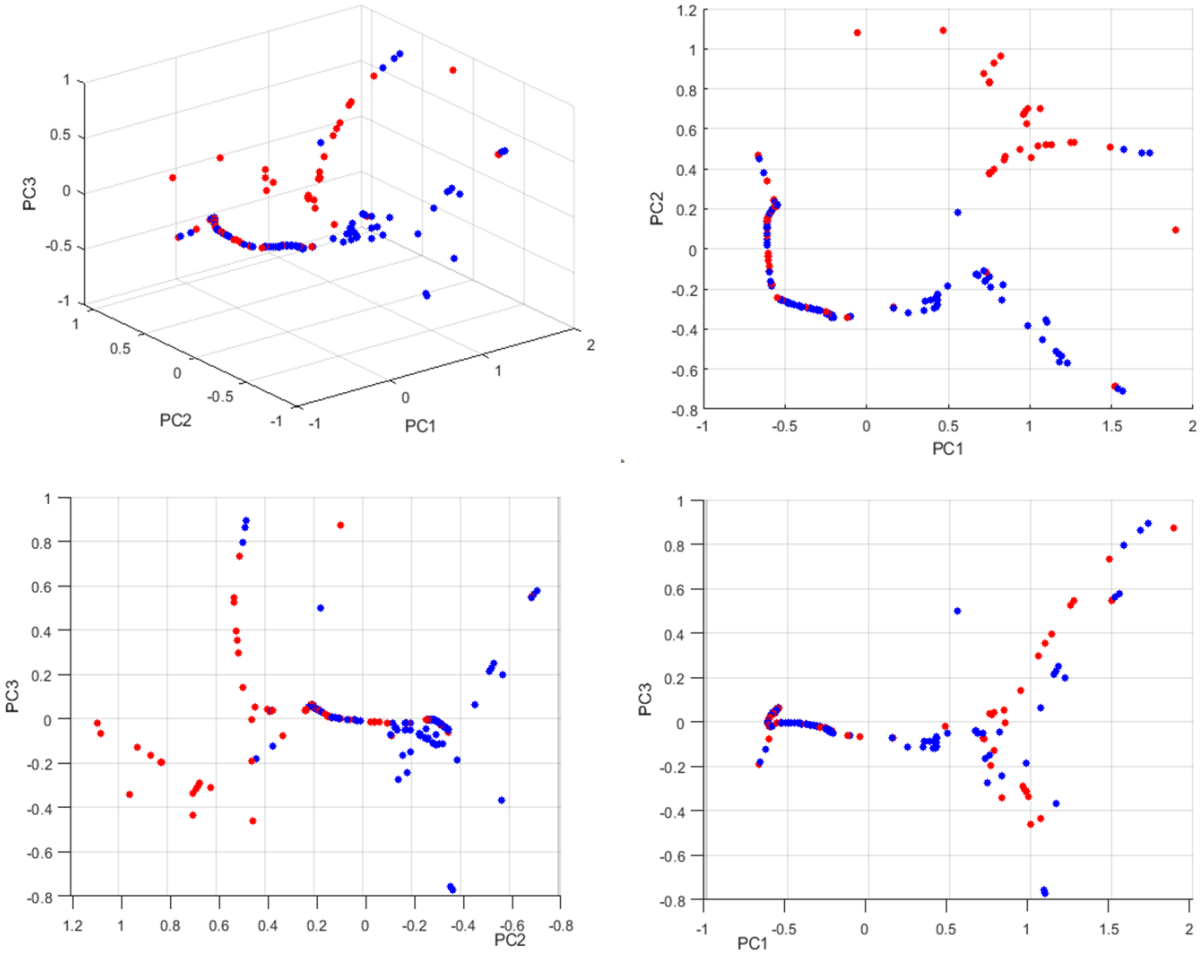


Figure 5-3: 3-dimensional and 2-dimensional visualisations of the calcification in the CBIS-DDSM training dataset, after the FIN and multidimensional scaling process. Note that Red is "Malignant"; Blue is "Benign"

Training data	Benign	Malignant	Consists mainly of…
Cluster 1	1	70	Malignant
Cluster 2	116	98	Unsure
Cluster 3	111	18	Benign
Cluster 4	13	5	Benign
Cluster 5	131	32	Benign

Table 5-2: Cluster distribution of the cases in the calcification subset of the CBIS-DDSM training dataset

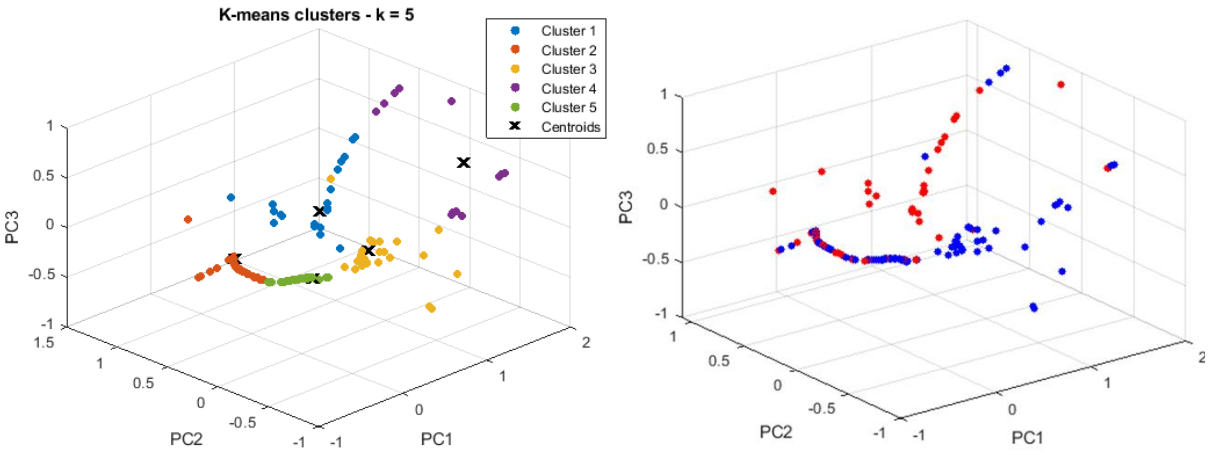


Figure 5-4: The training k-means clustering visualisation (left) and the embedding with MDS (right) side-by-side, for comparison. For the MDS on the right: ***Red is "Malignant"; Blue is "Benign"***

Figure 5-7 (left) shows the distribution of both the BIRADS scores and pathology across each trained cluster. Cluster 2 (orange), the cluster for which a main categorisation cannot be reached, contains a large number of cases that hold a BIRADS score of 4. At the same point within the embedding with MDS visualisation in Figure 5-4 (right), a mix of malignant and benign cases can be seen. This is similarly the case with cluster 5 (green), which resides close to cluster 2 within the embedding.

5.1.5.2 'Patient like me' approach

In this section, the application has been extended to the test cases.

Figure 5-5 shows 2-dimensional and 3-dimensional visualisations of the test cases projected onto the trained embedding, without clustering. Table 5-3 denotes the distribution of the cluster assignments of the test cases, upon the trained model. It can similarly be noted for the test cases, that cluster 1 consists mainly of malignant cases. Clusters 3, 4 and 5 consist mainly of benign cases, and the mix of both benign and malignant cases means that a categorisation here cannot be reached. This leads to the 'patient like me' approach where cases that did not inform the training of the model are projected onto the same embedding.

Figure 5-6 contains the test cases projected onto the k-means embedding. Figure 5-7 (right) denotes the distribution of the BIRADS scores and the pathology labels of the cases across the assigned clusters of the test cases. Similar to the training cases, it

can be seen that a majority of cases that hold a BIRADS score of 4 are within cluster 2.

The test cases projected onto the trained cluster embedding is shown in Figure S3.

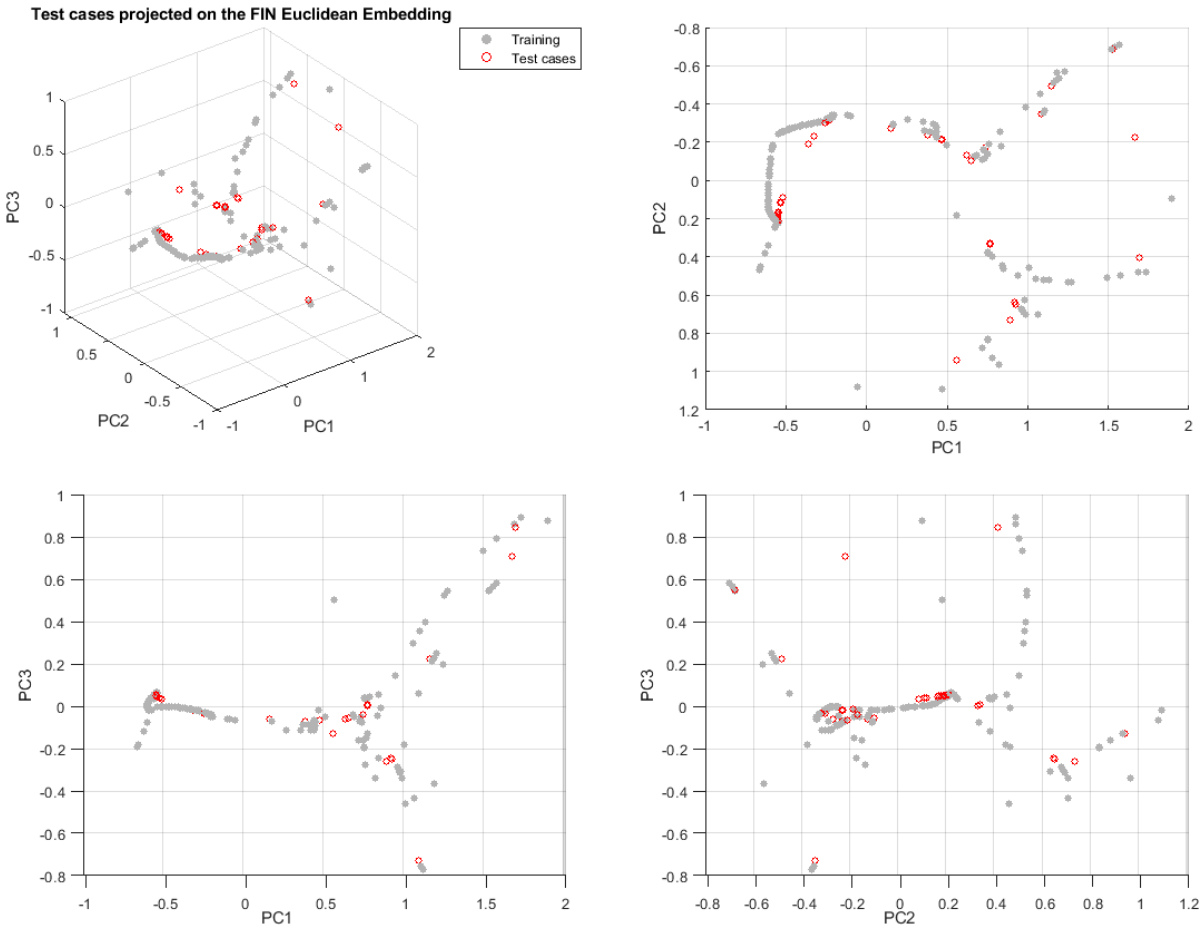


Figure 5-5: 3-dimensional and 2-dimensional visualisations of the test cases projected into the trained embedding. Note that red points are the new test cases.

Test data	Benign	Malignant
Cluster 1	4	12
Cluster 2	31	24
Cluster 3	17	4
Cluster 4	2	2
Cluster 5	21	7

Table 5-3: Cluster distribution of the cases in the calcification subset of the CBIS-DDSM test dataset

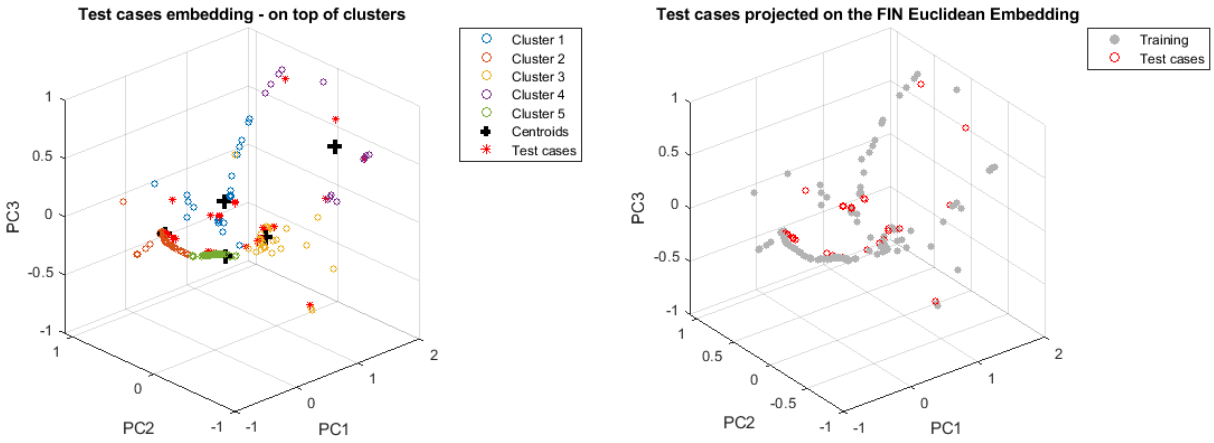


Figure 5-6: The projected test cases upon the k-means clustering visualisation (left) and the embedding with MDS side-by-side, for comparison.

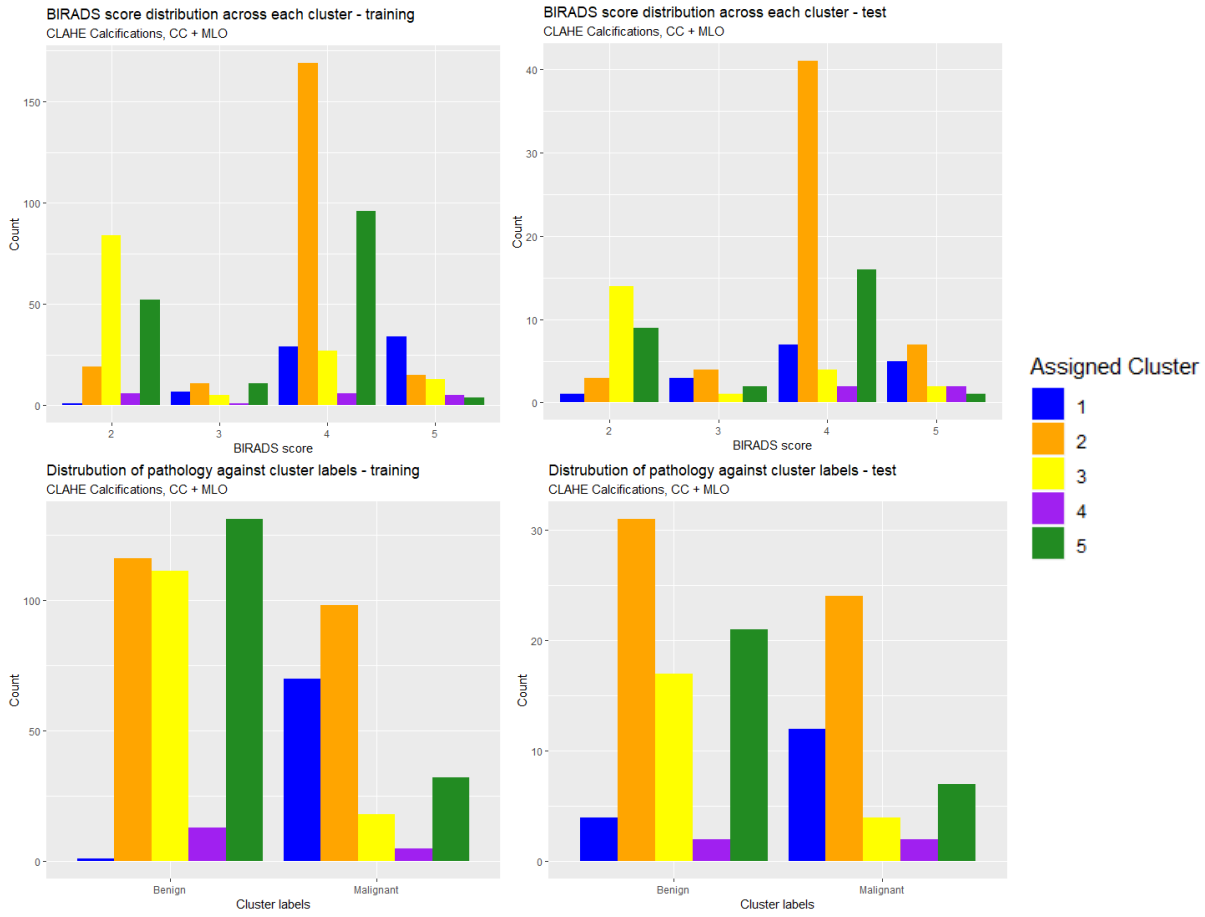


Figure 5-7: Distribution of BIRADS scores and pathology labels across clusters for the training (left) and test (right) sets of the data. The colours of each bar are similarly coloured for the cluster plots throughout the report.

5.1.6 Discussion

5.1.6.1 Latent space model created using breast cancer data – CBIS-DDSM

By projecting cases into the latent space model using the proposed methodology, it is intended that those cases with similar characteristics will be grouped together

within one part of the embedding, separately from cases with different characteristics, which will also similarly group. To this end, by creating a model using statistical texture features, of given ROI's of breast lesions, it is aimed that a minimal set of these features will effectively classify and group using the MLP.

The results of the MLP classification detailed in Table 5-1 indicate that the defined classification is reasonable. The nested cross-validation process assessed various combinations of the hyperparameter tuning process, with the best result presented. Using the Random Forest algorithm with feature importance for feature selection, interpretability is upheld which is in demand from clinicians in machine learning applications. Overfitting has been tackled by using a feature selection process to give a minimal number of features to the model as well as early stopping. This study looks at the representation of the model and so it was seen as acceptable to implement mitigation measures to this extent.

The results obtained using the training data (Table 5-2, Figure 5-4 and Figure 5-7) show that the proposed methodology can separate lesions holding similar characteristics. For example, training cluster 2 contains a large mix of benign and malignant cases. However, cluster 2 also contains a substantial number of cases which hold a BIRADS score of 4. This score denotes a "suspicious abnormality" – these cases should have a biopsy performed where possible. It is known that, if a BIRADS score of 4 is given, these cases can turn out later to be either a benign or malignant case, which follows the BIRADS score scale. In a similar vein, for training cluster 1 which contains a majority of malignant cases, the majority of these cases held BIRADS scores of 4 and 5. This is not perfect – the BIRADS score was not the outcome label for which the classifier was trained – however this extra information can provide valuable insight into the given cases.

This also follows with the test data – in Figure 5-5, the test cases are projected into the same embedding space as the training cases. It is expected that cases with similar characteristics 'land' within the training clusters/groupings that share the same characteristics as these new cases. Following this as shown in Figure 5-7, many cases that hold a BIRADS score of 4, are assigned to cluster 2. A majority of the test cases assigned to cluster 1 are malignant, holding BIRADS scores of 4 and 5.

5.1.6.2 'Patient like me' approach

It is intended that this embedding can be utilised as part of a decision support system for clinicians and radiologists in the diagnostic process of a patient's treatment care plan, which could be performed without the need for a pathological label. Where the relevant statistical texture features of the breast lesion can be taken from a mammogram, the embedding can project the case onto the relevant section of the embedding. Where a case is assigned to cluster 1 for example, the case likely exhibits similarities to the cases that are within the embedding and is likely to be a more severe case. This case should then be checked as a higher priority. If the case lands within clusters 3 or 5, it will share characteristics with other benign cases and the required action can be taken.

To further display the 'patient like me' approach, an example is provided in Figure 5-8. This shows that cases A and B that are correctly assigned to clusters that contain mainly malignant and benign cases, respectively, alongside their BIRADS score. Cases C and D were assigned incorrectly to clusters. Interestingly, both of these misclassifications hold BIRADS scores of 4; however, where this is the case, if as practice recommends, a needle biopsy is taken, and a pathology label may then be assigned. In the case of a patient's diagnostic care plan this may be seen as useful, to consider who to treat first.

It can be considered that to further improve the results and therefore the training of the FIN and MDS embedding, a better data representation must be taken. In the case of this work, the ROI's of the breast were of various different sizes – because statistical texture features were taken, the sizes of the images did not have to be amended or standardised, which can be seen as a positive against other fashionable algorithms; for instance, any information about the size of the images are not lost. However, this may explain the less competitive results attained here, than what is available in the literature.

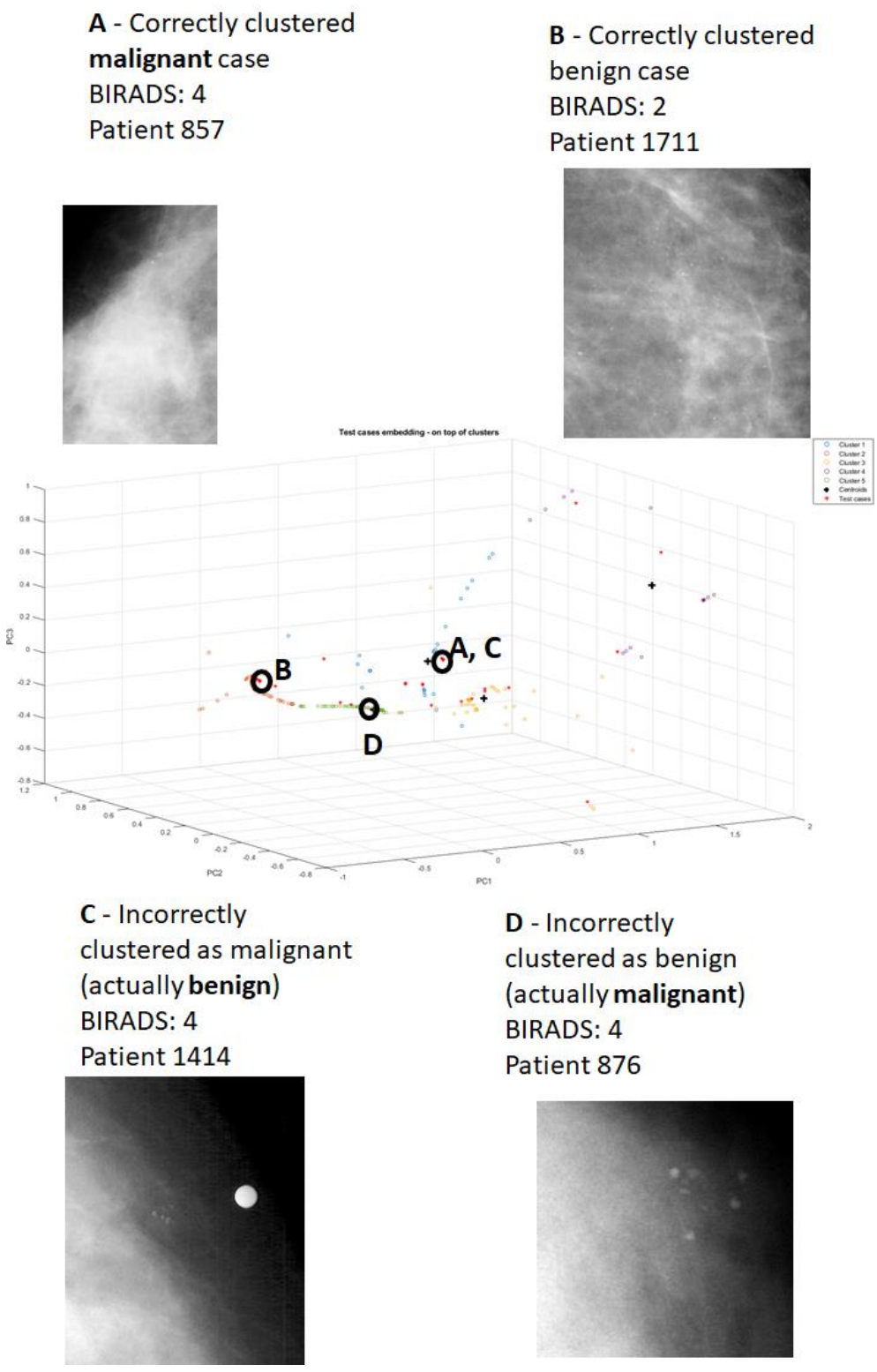


Figure 5-8: 'Patient like me' approach example - cases that are studied are circled in the test cases cluster embedding plot (note: there is overlap). These test cases reflect a new case that the model did not 'learn' and was not trained on.

5.2 Applying the FIN methodology to CNN features from a deep learning model

5.2.1 Introduction

This section builds on the previous presentation of applying the FIN through the creation of a robust representation of a breast cancer classifier, by extending the workflow into a strong deep neural network classifier. This study also uses the CBIS-DDSM dataset, although involves the use of a trained CNN model from the literature [100], [126]. Through using the CNN as a feature extractor and the weights and biases from the penultimate layer of the model, the FIN process shown in the previous section is applied to show the class separation of a deep learning model in the latent space.

The proposed methodology is like the previous section; however, this work studies features extracted from a CNN model and curates an MLP using this as the input data, rather than hand-crafted statistical texture features. This is done as to not ignore a powerful part of the classification literature, including strong results attained with the dataset [19], [100], [107], [127]–[129]. This work aims to add extra interpretability. Following the same process of constructing an FI network using probability density estimates but over five classes, the aim of this work is to "open the black box" of a deep learning model to exploit its predictive capabilities using this robust process.

The objective of this section is to visualise the process of a deep learning model (and not to affect the predictive side in any way). These more modern models have become known as one of the most difficult to interpret and this work attempts to add a layer of interpretability to the process. One may view this as a "post-processing" technique, which does not affect the classification process but shows the predictions in the latent space. This section will not study a patient-like-me approach although a similar process could be applied.

5.2.2 Methods

The methodology is shown in Figure 5-9 and described in this section of the chapter.

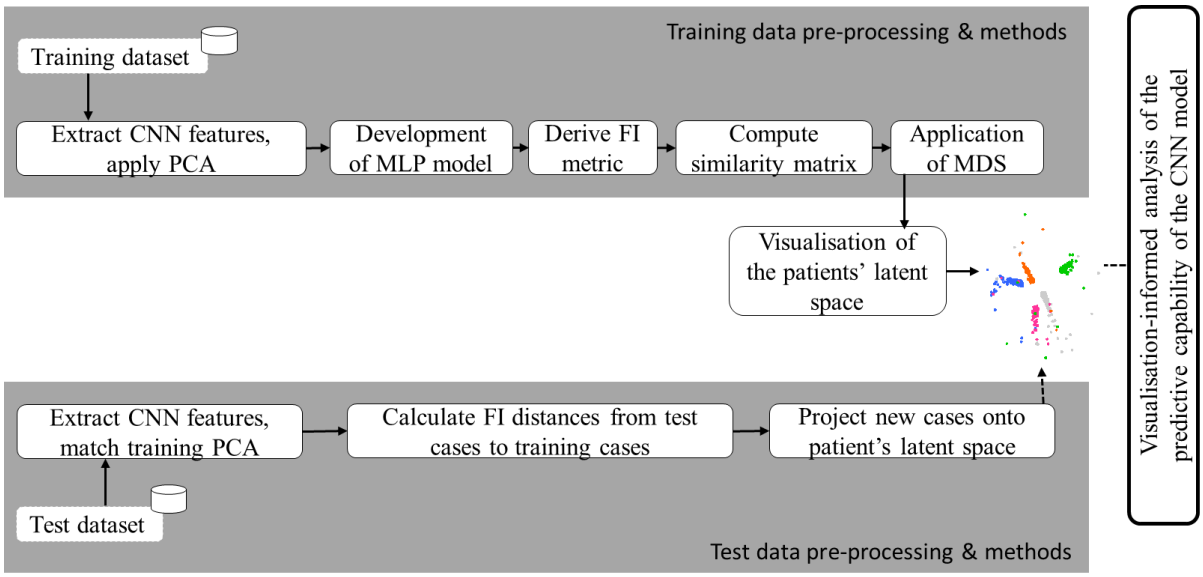


Figure 5-9: Methodology: Utilising the FIN methodology to create a visualisation of the latent space of cancer patients, exploiting the predictive capabilities of a CNN model.

5.2.2.1 Extracting CNN features

To study a CNN classifier as proposed, this work required a strongly performing deep learning model. The Resnet50-based model [130] is a 5-class patch classifier, trained on patches of mammograms around the lesion. Using the CBIS-DDSM dataset, the five classes are: calcification malignant, calcification benign, mass malignant, mass benign and (image) background. It is used later in Shen's work to develop a new classifier through transfer learning to inform a 2-class full image model, discriminating between the presence of cancer or none. A description of any data pre-processing, how the patches were attained and the curation of the five classes is described in [100].

This method is computationally expensive and so to reduce the extracted 2048 extracted CNN features, PCA has been used for dimensionality reduction. Out of the 2048 CNN features, 1363 principal components are kept after applying PCA, or 66.6% of the original number of features, capturing 90% of the variance. This CNN is used as a feature extractor in this work – the outputs of the penultimate layer are

extracted and used as the dataset for this work. Table 5-4 shows the size of the dataset for this study.

Label	Training	Test
Background	476	376
Calcification Benign	206	152
Calcification Malignant	110	91
Mass Benign	128	97
Mass Malignant	126	86

Table 5-4: Distribution of labels in this study.

A description of the aspects of the FIN methodology in this work – deriving the FI metric, computation of the similarity matrix and application of MDS leading to projecting new unseen test cases into the embedding – is described in the previous chapter.

5.2.2.2 Development of MLP model

Using the CNN features as the dataset for this work, an MLP was developed to discriminate between the five classes. The MLP was initialised with random weights, one hidden layer of 30 nodes and an output layer, a learning rate of 0.01 and a momentum of 0.9. Weight decay was implemented at a rate of 0.2. As a five-class classifier, the activation function was soft-max.

5.2.2.3 Using the FIN visualisations to analyse the predictive capability of the model

Firstly, the training features are used to derive the FI metric leading to the calculation of the similarity matrix. The application of MDS leads to the visualisation of each training case in the latent space. Separately the test set features are extracted and the FI distances of each testing point from each training point are calculated, which are then projected into the trained latent space. This replicates the process of training and then applying test cases to a machine learning model; the training of the model sets the foundations for applying the test cases.

To assess the capabilities of the CNN model using the Fisher visualisations, groupings shown in the FIN mapping will be analysed including an analysis on correctly classified test cases and their appearance on the visualisations. The

intention is to build on the process – first, showing the PCA analysis before applying the FIN methodology, then to show training cases only and finally to apply the test cases to the trained embedding.

5.2.3 Results

5.2.3.1 Model results

Although the classification results are not the focus of this work, they have been included to structure the discussion of utilising the visualisation to assess the capabilities of the model. Table 5-5 shows the confusion matrix for the training set, and Table 5-6 shows the confusion matrix for the test set. Overall, the model attains 72% testing accuracy and 98% training accuracy.

Training set confusion matrix (98% acc. overall)		**Predicted**				
		Background	**Calcification Benign**	**Calcification Malignant**	**Mass Benign**	**Mass Malignant**
Actual	**Background**	470 (98.7%)	2	0	2	2
	Calcification Benign	2	202 (98.1%)	2	0	0
	Calcification Malignant	2	3	105 (95.4%)	0	0
	Mass Benign	3	2	1	122 (95.3%)	0
	Mass Malignant	2	0	0	0	124 (98.4%)

Table 5-5: Confusion matrix of the CNN classifier for the training set. Percentages of the diagonal of the confusion matrix (TP/TN) presented to show correctly classified proportion in each class.

Test set confusion matrix (72% acc. overall)		Predicted				
		Background	**Calcification Benign**	**Calcification Malignant**	**Mass Benign**	**Mass Malignant**
Actual	**Background**	337 (89.6%)	14	7	11	7
	Calcification Benign	46	86 (56.6%)	14	2	4
	Calcification Malignant	12	24	42 (45.2%)	3	10
	Mass Benign	24	2	1	55 (56.7%)	15
	Mass Malignant	9	1	9	10	57 (66.3%)

Table 5-6: Confusion matrix of the CNN classifier for the test set. Percentages of the diagonal of the confusion matrix (TP/TN) presented to show correctly classified proportion in each class.

5.2.3.2 Before the FIN – PCA analysis

PCA analysis on the extracted features has been used to visualise the separation of the classes without applying the FIN process, as well as for dimensionality reduction before the process. Figure 5-10 (a) visualises the training data and Figure 5-10 (b) visualises both the training and test data when applying PCA on the 5-class classifier, before applying the FIN methodology and using the true (originally assigned) labels. Although some grouping can be seen across both training and test, the mixing of the classes is apparent. All classes overlap in the centre of the visualisation, and both malignant classes are deeply mixed. There is some improvement with the benign classes however mixing still occurs.

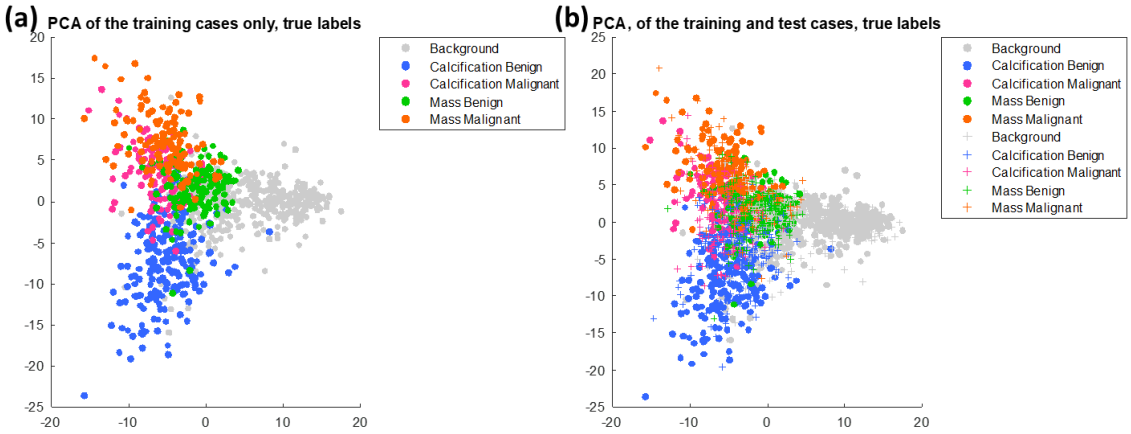


Figure 5-10: PCA visualisations of the data, using the originally assigned (true) labels - (a) only the training data; (b) training cases as spots, testing cases as crosses.

5.2.3.3 FIN visualisations of the training cases

Figure 5-11 (a) shows the visualisation using the FIN methodology of the classifier against the true labels, from which some limited mixing can be seen. This is information that we know from the data. Figure 5-11 (b) shows the visualisation but for how the model predicted for the training cases. This is an expected result and gives a view to how the classification process of the CNN model works. In Figure 5-11 (b) the classes are very well separated with most of the mixing occurring between the background and the calcification malignant classes. This figure can be used to assess how well the FIN representation reflects the MLP classifier. Through comparison of the visualisation with the true labels applied and the MLP predictions applied, the FIN representation looks to be suitable and reflects well. The MLP structure is well reflected in this study and can be seen as a reliable mapping of the training cases, for which test cases will be projected onto at a later stage.

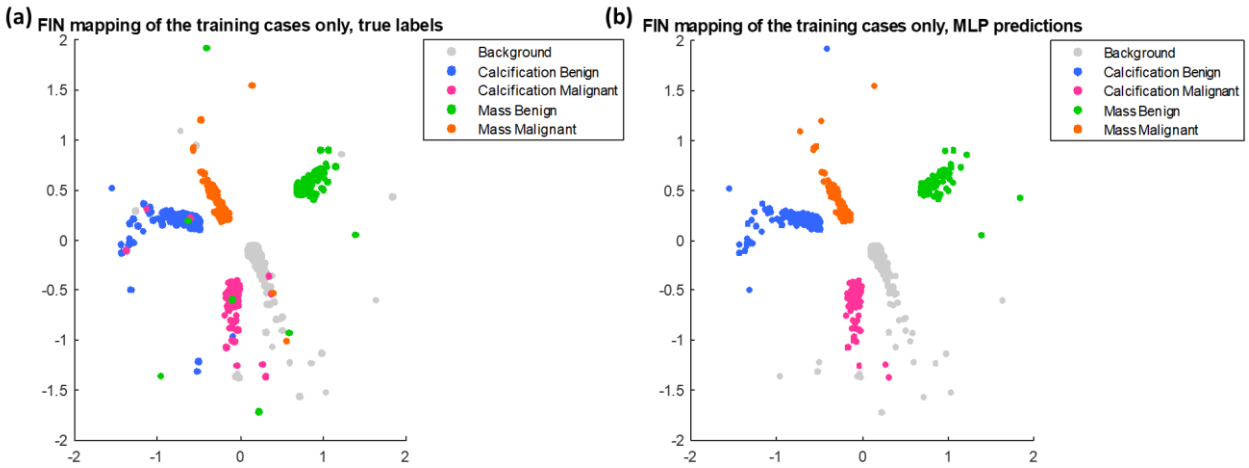


Figure 5-11: Visualisation of the training cases in the latent space: (a) against the true labels; (b) against the MLP prediction labels

5.2.3.4 Projecting the test cases onto the trained embedding

Figure 5-12 shows the correctly classified test cases of each class highlighted as a black marker on top of the more transparent training cases. These are the cases that have been classified correctly. For the test subset of the data, the background class contains the most correctly classified cases, with an accuracy of 89.6%. It is possible to see some of these correctly classified cases slightly spread out although these appear to be concentrated around the benign classes. The next best performing is the mass malignant class, with 66.3% of those test cases correctly classified.

The visualisation in Figure 5-12 shows that the correctly classified cases 'land' within the well-defined class groups Not all cases are correctly classified. A 'patient like me' analysis has looked at some of these incorrectly classified cases and is discussed further.

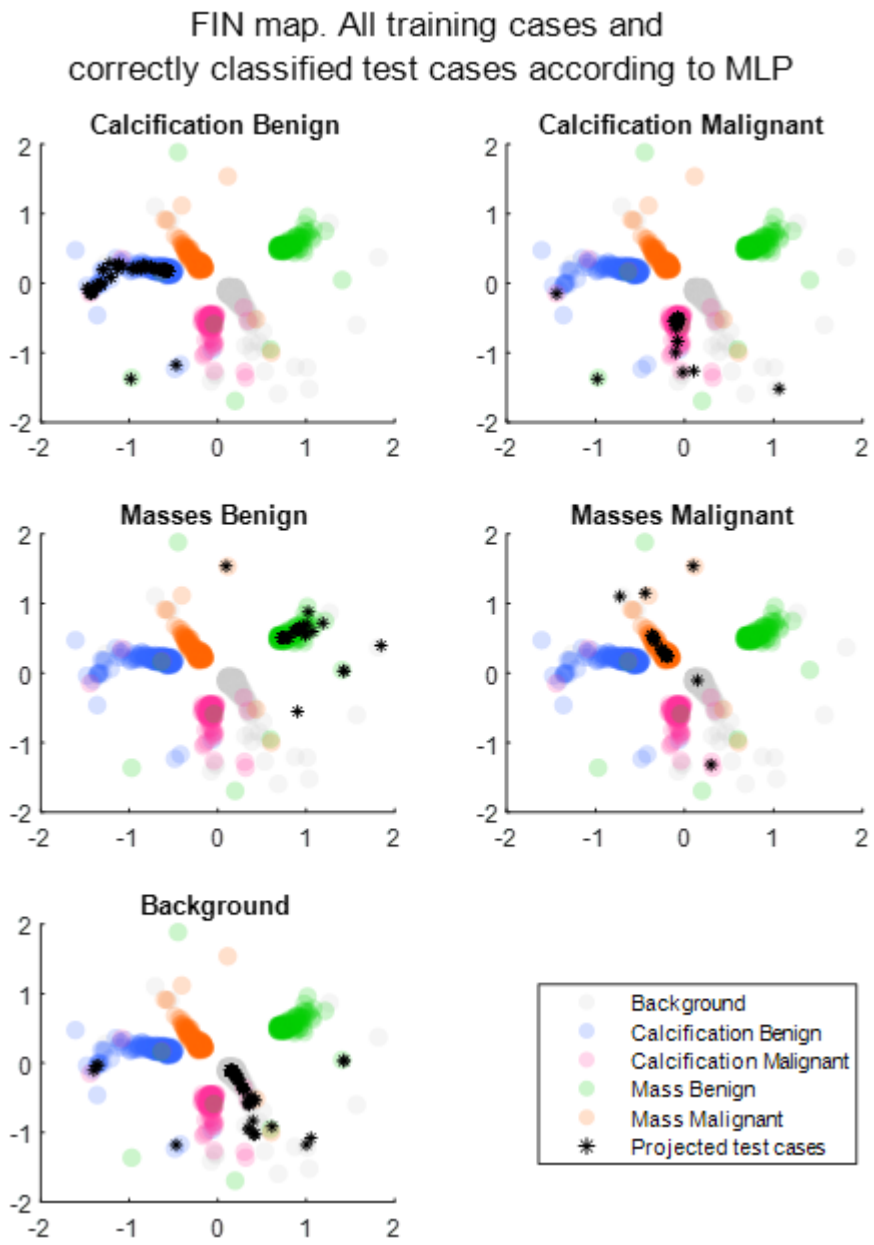


Figure 5-12: Correctly classified test cases (black stars) projected into the trained embedding.

5.2.3.5 'Patient like me' approach

Similar to the previous section, it is proposed that these trained embeddings can be utilised as part of a decision support system for clinicians and radiologists in the diagnostic process of a patient's treatment care plan, which could be performed without the need for a pathological label. Through the extraction of the CNN features, using the same model, can lead to the new case being projected onto the embedding. This can provide new insight into the case. In this 'patient like me' analysis, metadata unseen by the classifier has been used to extract further insights into why cases have been (mis-)classified in the way they have.

Figure 5-13 shows a ‘patient like me’ analysis of four correctly classified cases with pathological labels. For a given test case in the trained embedding, its location has been highlighted along with the ROI patch classified, information and metadata about the given case. Both malignant cases – patients 534 and 146 – hold BIRADS scores of 5 and high subtlety scores.

Figure 5-14 shows a similar ‘patient like me’ analysis, but for five cases that were misclassified. Importantly for these cases, the metadata can provide insight into why the cases were misclassified. For example, for both patient 1569 – a calcification benign case incorrectly classified as calcification malignant - and patient 420 – a mass malignant case incorrectly classified as mass benign - the associated metadata shows that the BIRADS score assigned is 4 (BIRADS was not a feature in the classification, and is denoted from studying a mammogram alone, pre-pathology). This score is defined as “suspicious of malignancy”, with a score of 3 or 5 leaning towards benign or malignant, respectively. This could support the fact that the cases were misclassified.

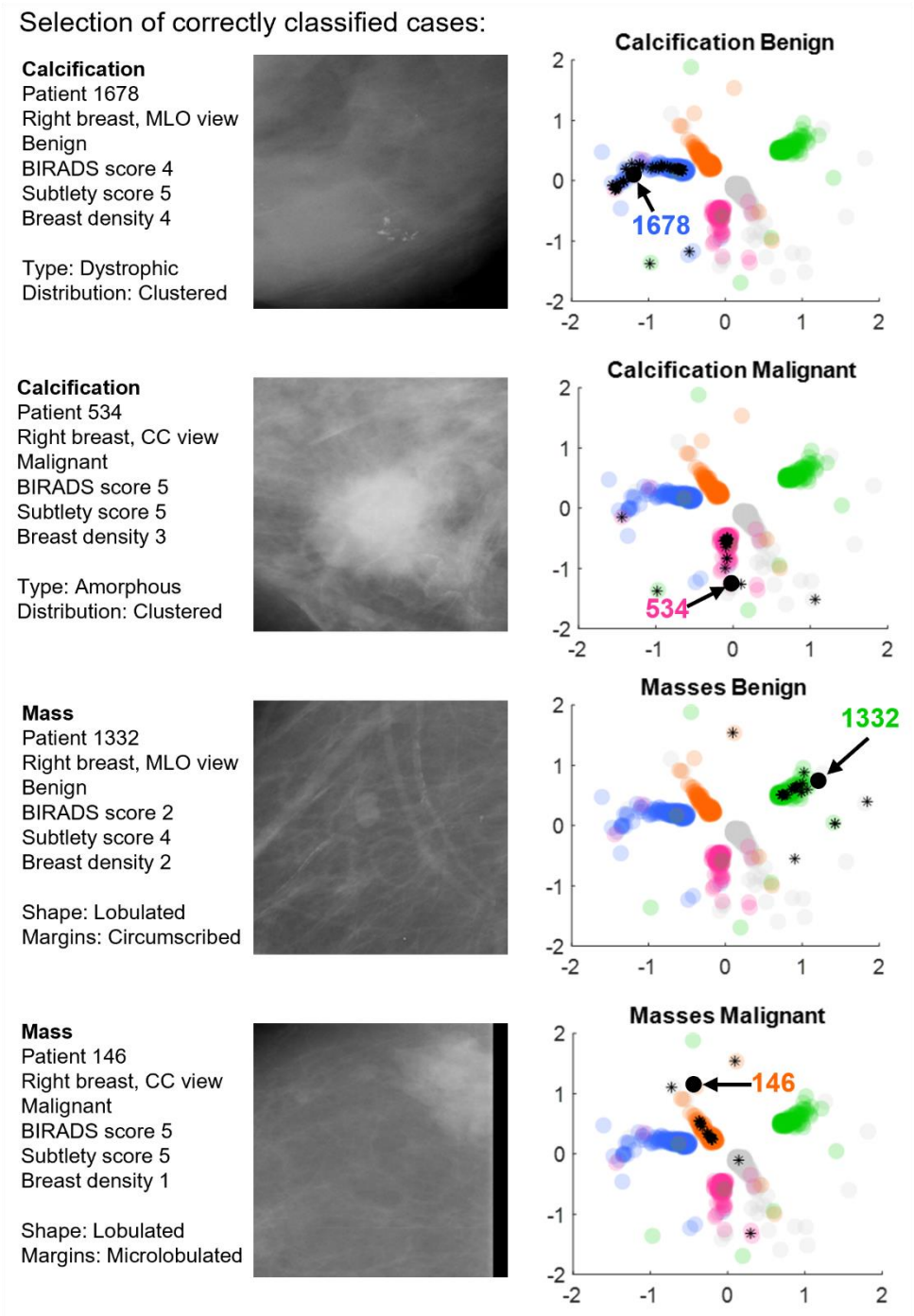


Figure 5-13: 'Patient like me' analysis of four cases correctly classified with pathological labels. ***Patient 1678*** *is a correctly classified Calcification Benign case,* ***Patient 534*** *as calcification malignant,* ***Patient 1332*** *as mass benign,* ***Patient 146*** *as mass malignant. Black stars are correctly classified test cases, black circle is the given example.*

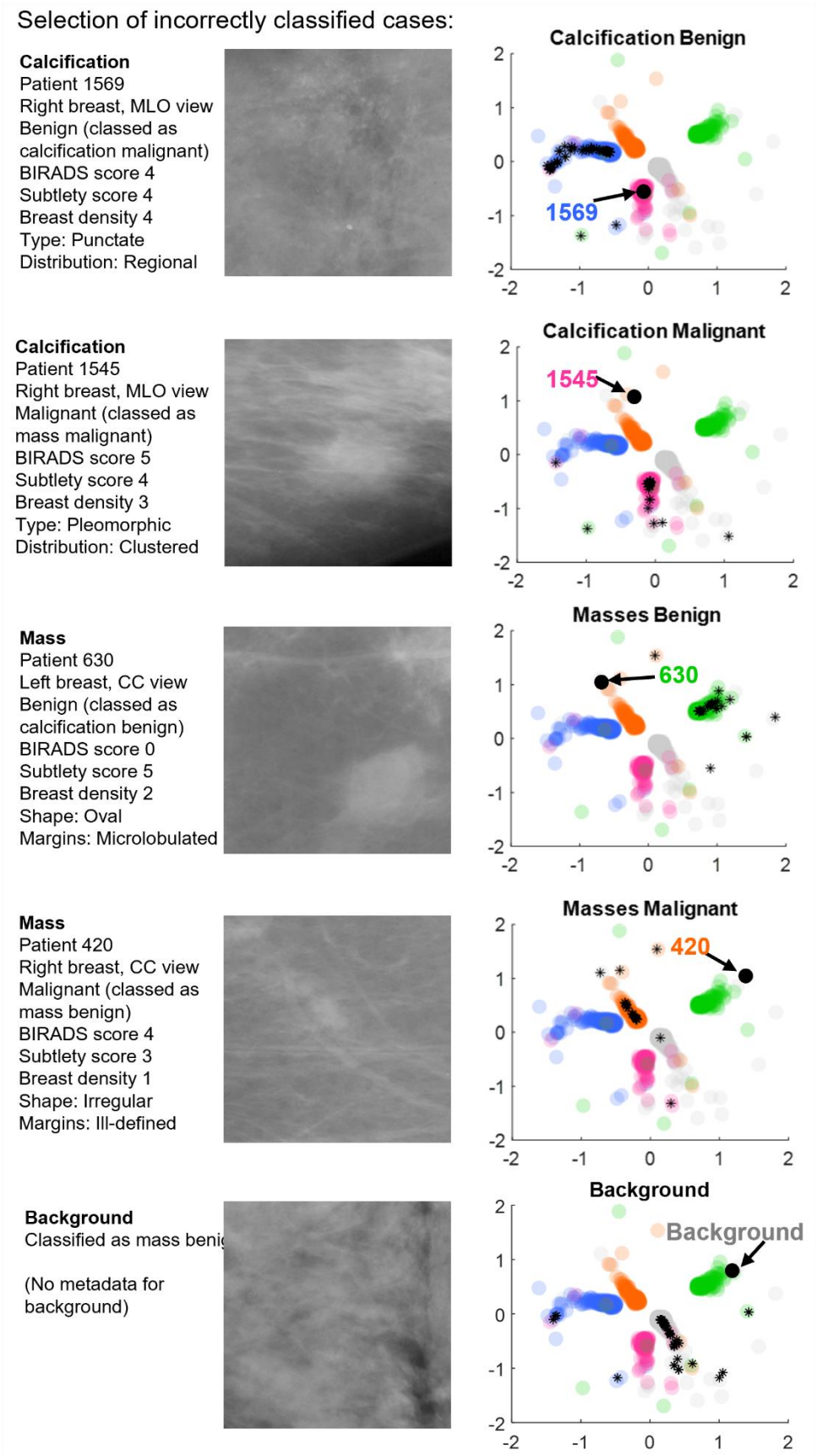


*Figure 5-14: 'Patient like me' analysis for five misclassified cases. **Patient 1569** is a misclassified benign calcification (classed as malignant calcification), **patient 1545** is a misclassified malignant calcification (classed as malignant mass), **patient 630** is a misclassified benign mass (classed as benign calcification), **patient 420** is a*

misclassified malignant mass (classed as benign mass), a background patch is misclassified as a benign mass. Black stars are the correctly classified test cases in each of the four groups -added for reference- and the selected cases are represented with black circles.

5.2.4 Discussion

5.2.4.1 Visualisation-informed analysis of the CNN model

The trained embeddings shown in Figure 5-11 provide an indication of separation using the FIN as a "post-processing" step. Alongside the MLP training accuracy of 98% and each of the five classes in the training set attaining their own respective accuracies over 95% the FIN methodology has been able to represent these CNN features strongly. This provided a good embedding to progress with allowing for the projection of the test cases to continue.

As denoted in Table 5-6 for all four lesion classes, most of the misclassified cases land in the background class, which is evident in the visualisation. Throughout each class, however, the misclassifications generally spread throughout the other classes. In this work, it is not the quality of the classification that is of concern, but what the visualisations can tell us about the CNN features and how the classifier has "thought".

Although developing the classifier itself was not the focus of this work, some short commentary is presented. This classifier is a 5-class patch classifier, as it was trained to be in Shen's work. The aim of Shen's work was to curate the patch classifier and apply transfer learning to develop it into a classifier which discriminates whole images into two classes – whether cancer is present or not. This model could be seen as an initial "triage" model, where any given mammogram patch can be inputted to give an initial idea.

This work has reviewed a 5-class classifier, which includes calcifications, masses, and mammogram background. However, calcifications and masses are different by their very nature and are both assessed in this classifier. In the previous section of this chapter the FIN is applied to calcifications only to assess this viewpoint. Furthermore, other chapters in the thesis review the dataset but separate the masses and calcifications before conducting the study.

There is sign of overfitting in the trained CNN. Although this is known as a typical problem for deep learning models, this work does not attempt to solve the overfitting problem but instead aims to visualise a CNN's classification process to show how this methodology could be used in a real-world scenario, as a proof of principle. Due to technical and computational restrictions, training the "best model" was outside the scope of this thesis.

5.2.4.2 Why is this useful in a clinical setting?

In this work, it has been possible to study a given case that has been classified using CNN features and involve the metadata, which has been unseen to the classifier, to gain extra insight into the finding. It is proposed that this can reflect clinical thinking and decision-making. With further knowledge from a good classifier and visualisation, this tool can be another aspect in the process.

Where a new test case lands within the trained embedding, it is intended that this new case will be like those around it. For example, the correctly classified cases (as shown in Figure 5-13) can be used as examples of correctly labelled and classified cases. Studying the associated training cases within the embedding can strengthen this case and assist with confidence in the tool. Whereas incorrectly classified cases (as shown in Figure 5-14) require further work. For example, cases E and F were incorrectly classified, the type of lesion correctly but the classification of pathology not. The BIRADS score in this case was "suspicious of malignancy", the borderline score in this classification, and has given some idea as to why this classification is wrong.

It is intended that the FIN methodology proposed throughout this chapter can be seen as working "hand-in-hand" with clinicians and radiologists, using this extra information to assist in a 'patient like me' approach to attain further insights into the workings of the classifier and visualisations. To improve this further, future work could consider the use of active learning [131] to continuously improve the embedding. In theory, this could be a triaging tool, particularly for the work in the previous section on using statistical texture features. Before any pathology is taken, a case can be viewed to see where it lands, as a case may exhibit such characteristics. Then, where a pathology is taken, this information can be included

and the methodology updated to improve the model, leading to a better ‘patient like me’ approach for future patients.

5.3 Conclusions

The clinical detection and diagnosis of cancers is an important and delicate challenge for multidisciplinary healthcare teams in patient care. Decisions must be based on evidence gathered from various techniques. In breast cancer, the use of mammography is standard, although difficulty in its use remains prevalent in clinical practice.

The first part of this chapter looked to enhance the understanding of a curated MLP to strengthen the detection of breast cancer using mammograms, the FIN has been utilised, alongside MDS. This has provided an informed embedding, leading to a 'patient-like-me' approach, aiding in the detection of breast cancer in medical imaging.

The aim of the second part of this chapter was to exploit the predictive capabilities of a deep learning model. The Fisher Information methodology has been shown to "open the black box" of the CNN classification process in this chapter. The Fisher Information metric has allowed for visualising the usual classification process, proving a solution that is clearly defined and statistically strong to show the "thinking" of the CNN model.

A standalone CNN model can provide probabilistic output albeit with little explanation as to why. This model can treat each case individually, while still utilising a powerful machine learning model in the literature. This process has also given a visual representation of the predictive capability of the deep learning model which exploits the predictive capabilities in a robust mathematical manner.

The use of the FIN methodology defined in the previous chapter has been extended to show that a similar process can be applied to represent clinical data, classified through a CNN model which is traditionally difficult to interpret. Through this method, extensive analysis has taken place on assessing the predictive capabilities of the model, in particular the sensitivity and specificity. A Fisher-informed approach can add insight into the predictive capabilities and as described, can assist in the breast cancer diagnostic process.

6 Partial responses to enhance the interpretability of a breast cancer classifier

6.1 Introduction

A multilayer perceptron (MLP) and its partial responses have been studied in this chapter, to build an explanatory risk predictor of breast cancer malignancy. The aim is to enhance the understanding of a traditionally “black box” method, the MLP, by explaining the predictive capabilities and output of the model. It is intended that the predictions of the model will be explainable through the calculation of partial responses of individual variables, showing the effect of the outcome through individual variables.

Analysis of partial responses of an MLP model [132], [133] provides an interpretation of model predictors, aiming to “open the black box” of the methodology. The partial response of each variable can be visualised against its contribution to the logit, which leads to a key level of detail on how specific variables influence the outcome. Tabular data, of statistical texture features and associated metadata for breast lesions, has been used to construct a more transparent and interpretable model.

Although other machine learning techniqueshold inherent interpretability – such as decision trees and k-Nearest Neighbour algorithms – this study considers the stronger performance of MLPs and describes a view to “open the black box” of the classifier.

The objective of the work presented in this chapter is to study the effect of each variable of a breast cancer classifier and attempt to interpret the results. From this, insights about how each variable interacts with the logit and how it affects the outcome overall. Throughout this chapter, the partial response methodology is applied to both continuous and categorial variables. The workflow for this chapter is illustrated in Figure 6-1.

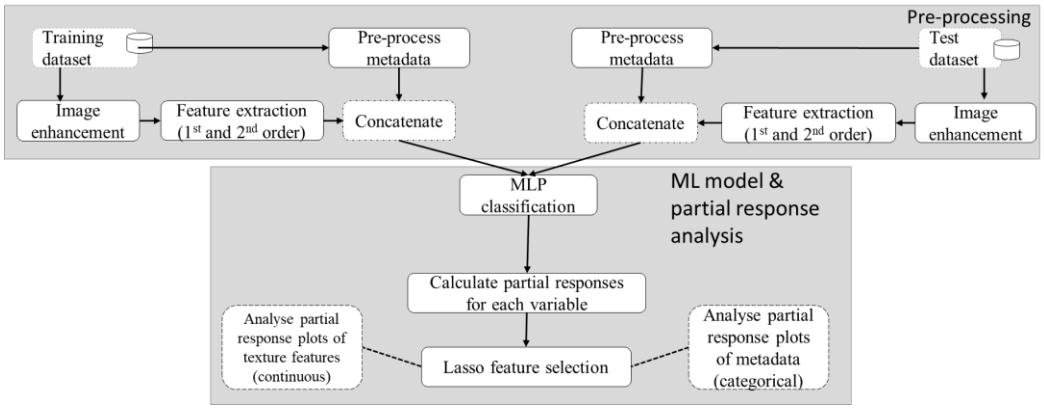


Figure 6-1: Workflow described throughout the partial response chapter.

6.2 Materials and methods

6.2.1 Data utilised from CBIS-DDSM

Due to their visual differences, the work is conducted separately on breast calcifications and breast masses with results presented for both lesion types. Two data subsets are used here – the statistical texture features and the associated metadata. These are described in Chapter 3.

6.2.2 Multilayer perceptron

This study will look at binary classifiers, one for breast masses and another for calcifications. To review the partial responses of a model, MLP classifiers will be used to discriminate between benign and malignant breast tumours. An MLP is a type of feed-forward neural network, fully connected by neurons. A strength of MLPs is the fact that nonlinear relationships in the data can be mapped. The sigmoid activation function will be the probability estimation evaluator.

6.2.3 Partial responses

Partial responses [134] attempt to explain the decision-making process behind the MLP. The response from an MLP when all but a specific variable being analysed are zero, is known as the partial response. Partial responses are calculated by providing one input at a time to an MLP, so that the contribution of each variable to the log of the response can be calculated.

After fitting the data with an MLP, the terms in the ANOVA decomposition are calculated, as low-order dependencies in the probability density function can be extracted with it. This decomposition comprises a finite number of terms up to interactions of dimension, d. Each term is orthogonal in a functional sense [135] and so may be regarded as independent inputs later.

$$\begin{aligned} & logit\big(P(C|x)\big) \\ & \quad \equiv \phi(0) \\ & \quad + \sum_i \phi_i(x_i) \\ & \quad + \sum_{i \neq j} \phi_{ij}(x_i, x_j) + \cdots + \sum_{i_1 \neq \cdots \neq i\_d} \phi_{i_1 \ldots i_d(x_{i_1}, \ldots, x_{i_d})} \end{aligned}$$

Equation 6-1

Each term from Equation 6-1 is computed from the logit of the MLPs output, $logit(P(C|x))$ which is the log-odds probability of class membership for a given input vector as such:

$$\phi(0) = logit\big(P(C|0)\big)$$

Equation 6-2

$$\phi_i(x_i) = logit\left(P\big(C\big|(0, \ldots x_i, \ldots, 0)\big)\right) - \phi(0)$$

Equation 6-3

$$\phi_{ij}(x_i, x_j) = logit\left(P\left(C\middle|\big(0, \ldots x_i, \ldots, x_j, \ldots, 0\big)\right)\right) - \phi(x_i) - \phi(x_j) - \phi(0)$$

Equation 6-4

Then, the logistic Lasso is applied for hard feature selection with l_1 regularisation. The inputs correspond to the partial responses as defined – where all other variables are set to zero and the response function of the MLP is recalculated for the modified input vector, deriving its logit.

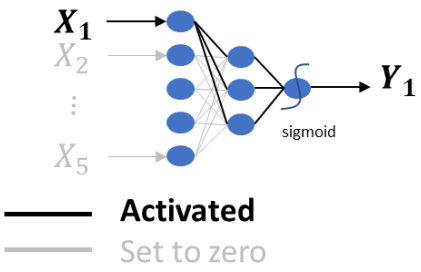
Finally, as the partial responses are obtained through an MLP output (from its weights), the feedforward structure of the Lasso with only the selected partial response functions is replicated using the original weights in the form of a general additive neural network.

To model the contribution to the logit using univariate terms, the model is defined as:

$$\log(Y) = \phi(0) + \sum_i \phi_{i_{(x_i)}} + \epsilon$$

Equation 6-5

Where $\phi(0)$ is the error that is calculated when all inputs are equal to zero, and $\phi_{i_{(x_i)}}$ represents the partial responses of a given variable, i (individually) and ϵ represents the higher-order terms. Figure 6-2 illustrates this within an MLP. In this work, a larger contribution to the logit denotes an increased risk of malignancy.



$$Y_1 = \theta(v_k) = \frac{1}{1 + e^{-av_k}}$$

Partial response:

$$\log(Y) = \phi(0) + \sum_i \phi_{i_{(x_i)}} + \epsilon$$

Figure 6-2: Partial response within an MLP, where only X_1 is activated and its partial response analysed.

6.2.4 Feature selection

The Lasso is a shrinkage method which reduces the number of variables selected in a dataset, known as the "Least Absolute Shrinkage and Selection Operator," using an L_1-penalisation term on the minimisation function:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}}\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p}|\beta_j|\}$$

Equation 6-6

Because of this penalisation term, the absolute values of the coefficients can reduce to zero, leading to a method of variable selection in this process.

Although all variables will have their own partial responses to the model, this chapter will look at the variables noted as important by the Lasso feature selection process. The results presented all use the Lasso parameter lambda which is within one standard error from the minimum. This is traditionally used to select the best model as it "acknowledges the fact that the risk curves are estimated with error, so

[this concept] errs on the side of parsimony" [136]. It chooses the simplest model for which the accuracy is comparable with the best model [137].

6.2.5 Evaluation of features

To evaluate the partial responses of the features, cases with a lower contribution to the logit are more likely to be benign, while cases with a larger contribution to the logit are more likely to be malignant. All plots and partial responses are calculated using the full dataset.

6.3 Results and Discussion

6.3.1 MLP setup

The MLPs for both masses and calcifications are separate but are set up the same, as follows. An input layer with 13 nodes, a hidden layer with 5 nodes and an output layer, sigmoid activation function. Dropout is applied to the input and hidden layer at rates of 0.4 and 0.3 respectively. The optimiser used is the Adam optimiser and binary cross-entropy is the loss function.

6.3.2 Masses

6.3.2.1 MLP setup and results

Table 6-1 shows the dataset split of the mass subset of the CBIS-DDSM dataset for this work. There were 130 statistical texture features and 16 categories of metadata available for the mass subset. The MLP setup was an Adam optimiser and binary cross-entropy loss. The MLP attained an AUC on the test set of 0.93, with a loss of 0.362.

Masses	**Training**	**Testing**	**TOTAL**
Benign	512	150	662
Malignant	502	84	586
TOTAL	1014	183	1197

Table 6-1: Dataset split for the mass subset.

6.3.2.2 Lasso feature selection

Table 6-2 shows the 7 texture features (5.4% of the 130 available) and the 7 metadata categories (43.75% of the 16 available) selected through the Lasso and

their lambda coefficients. These variables and their partial responses are analysed further. The test AUC of the Lasso model was 0.949.

Texture Features		Coefficient	Metadata		Coefficient
	(Intercept)	0.018	Shape	Irregular	0.8123
First order	Energy	0.4978		Lymph Node	0.0133
	Kurtosis	0.1431	Margins	Circumscribed	0.9512
	Maximum	0.5106		Ill Defined	1.6947
	Minimum	0.3281		Obscured	0.7927
Second order	GLCM Auto Correlation (angle 135)	0.7326		Spiculated	2.2151
	GLCM Inverse Variance (angle 135)	9.4121		Microlobulated	1.8543
	GLRLM SRHGLE (angle 45)	1.1261			

Table 6-2: Mass subset, Lasso selected features and their lambda coefficients

Figure 6-3 shows histograms of the frequency distributions and contribution to the logit of the continuous texture feature variables. Cases with higher values of kurtosis and minimum are more likely to be benign. Cases with higher values for energy, maximum, GLCM Auto Correlation, GLCM Inverse Variance and GLRLM SRHGLE are more likely to be malignant. In mammograms, normal fatty tissue is usually grey and tumours white. Energy is calculated as the square sum of each matrix element, with higher values suggesting more intensity variation. It could be suggested that a high value for energy - a large change in image intensity - is seen with a more severe mass within the ROI, allowing for some enhanced interpretability to clinicians. In practice, these cases may require further investigation.

Another example is a second-order feature (two pixels defining the GLCM feature), Auto Correlation. It measures the coarseness of an image where the higher the value, the greater the concentration of low grey values within the image. As seen in Figure 6-3, a lower value of GLCM Auto Correlation means the case is more likely to

be benign and more likely malignant when the value is higher. Here, the values of GLCM Auto Correlation will increase due to the change in texture of the image.

Figure 6-4 shows histograms for the Lasso selected categories with the red points representing the contribution to the logit of the given variable. In Figure 6-4 it is shown that when the mass shape is irregular, the contribution to the logit is higher (more likely to be malignant) than if it is not. Where the shape is recoded as a lymph node, the contribution to the logit is smaller (more likely to be benign). Bassett and Conner [92] state that an irregular shape suggests a greater likelihood of malignancy. This feature from the model follows clinical guidance.

Figure 6-4 also shows that if the margins are either ill defined, spiculated or microlobulated then according to the model, the risk of malignancy is higher than if the given mass lesion is not. This also follows clinical knowledge [92], [138].

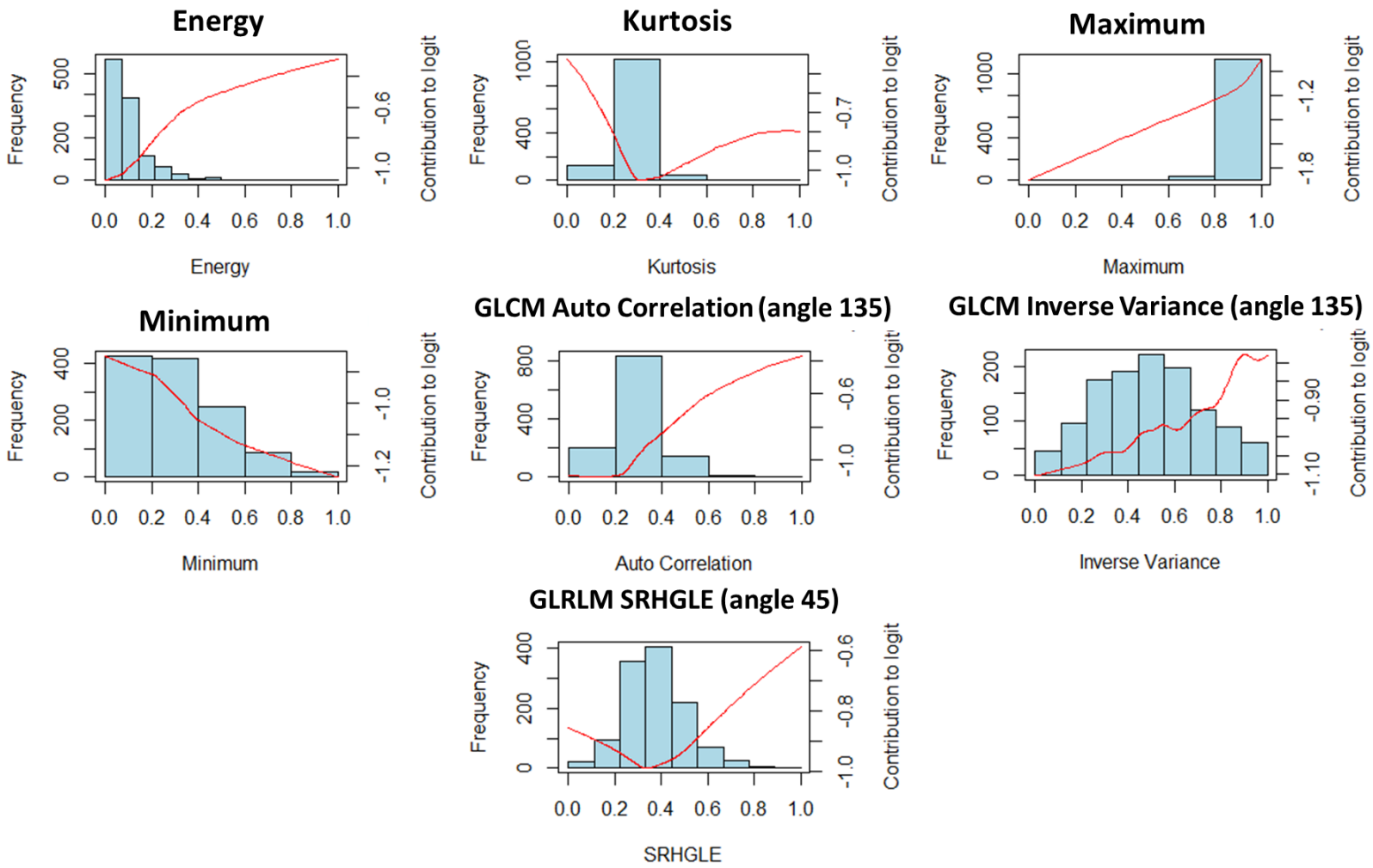


Figure 6-3: Mass subset: Partial responses (red line against y-axis on the right) for the continuous Lasso selected features (texture features

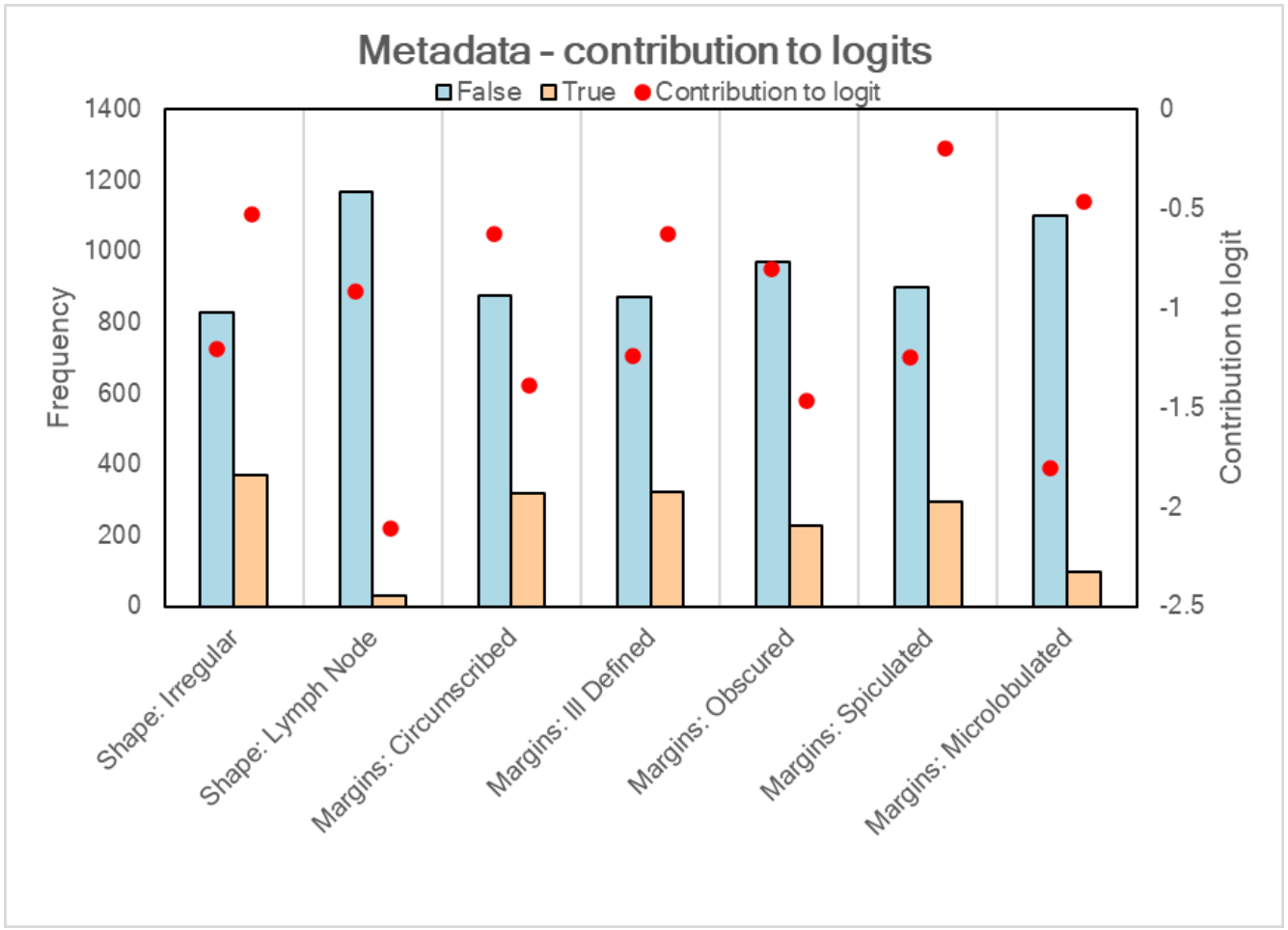


Figure 6-4: Mass subset: Partial responses (red points against y-axis on the right) for the categorical metadata. Note, "True" denotes the given metadata aspect is true (=1).

6.3.3 Calcifications

6.3.3.1 MLP setup and results

Table 6-3 shows the dataset split for the calcification subset of the CBIS-DDSM dataset for this work. The MLP setup was the same as for the mass subset. The MLP attained an AUC on the test set of 0.853, with a loss of 0.480.

Calcifications	**Training**	**Testing**	**TOTAL**
Benign	798	150	948
Malignant	418	84	502
TOTAL	1216	234	1450

Table 6-3: Dataset split for the calcification subset.

6.3.3.2 Lasso feature selection

Table 6-4 shows the 14 texture features and the 15 metadata categories selected through the Lasso and their lambda coefficients. These variables and their partial responses are analysed further. The test AUC of the Lasso model was 0.877. It should be noted that the "N/A" in Type and Distribution mean that a label for the Type or Distribution are "not applicable" rather than missing. In mammograms, normal fatty tissue is usually grey, and calcifications are traditionally white spots (in a similar way as masses are white).

Figure 6-5 shows the histograms for the Lasso selected texture features, and breast density with the red lines representing the contribution to the logit for the given feature. The breast density partial response increases (more likely malignant) as the breast density increases, dropping slightly as breast density reaches the highest category. This follows the literature on breast density being correlated with breast cancer [90], including the fact that younger women tend to have denser breasts and there is a correlation between age and breast cancer[3].

As shown in Figure 6-5 for the texture features energy, maximum, GLCM correlation (angle 0), GLCM entropy (angle 90), GLRLM GLN (angle 90) and GLRLM RLN (angle 90) as the values of these features increase, so does the likelihood of malignancy.

[3] Age was not available with the CBIS-DDSM dataset. It was available with the DDSM dataset however it was not possible to transfer this over retrospectively.

These are monotonic partial responses, where the feature is directly correlated with malignancy. GLN (grey level non-uniformity is lower if intensity values are alike, so in this case where intensity values differ the risk of malignancy increases. GLRLM RLN (run length non-uniformity) measures the similarity of the length of the runs throughout the image and is lower if the run lengths are alike and so at angle 90 as the run lengths (runs of similar pixel values) are less alike, the risk of malignancy increases.

For the texture features GLCM dissimilarity (angle 0), GLCM contrast (angle 90), GLCM difference entropy (angle 90), GLRLM LGLRE (angle 135) and GLRLM RLN (angle 135), as the value of the feature increases, the likelihood of malignancy decreases. These are also monotonic partial responses, where it is inversely correlated with malignancy. Difference Entropy measures the disorder related to the grey level difference distribution of an image and so as this increases, the likelihood of malignancy decreases.

V-shaped partial responses are shown for the texture features GLCM cShade (angle 90), GLCM sum entropy (angle 90) and GLRLM HGLRE (angle 135). This suggests that the chances of malignancy change as the values move away from the median value. In practice, this could suggest that values near the median require urgent review such as in usual clinical practice for lesions suspicious of malignancy.

Figure 6-6 shows the histograms for the Lasso selected categories with the red points representing the contribution to the logit of the given variable. In Figure 6-6 it is shown that when the calcification type is listed as amorphous, pleomorphic, or fine linear branching it is more likely to be malignant as the contribution to the logit is larger, albeit marginally. This is supported by Bassett and Conner [92]. Where the calcification type is noted as coarse, dystrophic, punctate, milk of calcium, large rodlike, ‘round and regular’ or vascular, it is more likely to be benign. This is also supported by the clinical literature [92], [139], [140].

Figure 6-6 also shows that if the distribution is listed as linear it is more likely benign, which follows the literature [92]. However, although the partial responses show that clustered distributions are more likely to be malignant, this does not

follow the literature as stated by Bassett and Conner [92]. It is worth noting that for the clustered distribution as shown in Figure 6-6, the gap between the partial responses for a case being benign or malignant and its contribution to the logit is small, so in this case it may be circumstantial.

There is a similar case for the diffusely scattered distribution for which the partial responses dictate that the presence of this category is more likely to be benign. This does not follow Bassett and Conner who state that this category means a case is more likely to be malignant.

Texture Features		Coefficient	Metadata		Coefficient
	(Intercept)	-0.4407		Breast Density	1.0226
1st order	Energy	3.0129	Type	N/A	0.1511
	Maximum	-0.0117		Amorphous	0.4932
2nd order GLCM	Correlation (angle 0)	-3.5451		Coarse	0.0113
	Dissimilarity (angle 0)	-2.0408		Dystrophic	0.1588
	cShade (angle 90)	1.6545		Pleomorphic	1.2131
	Contrast (angle 90)	5.0245		Punctate	0.2942
	Difference Entropy (angle 90)	1.1511		Milk of Calcium	0.2764
	Entropy (angle 90)	-1.0283		Fine Linear Branching	2.0384
	Sum Entropy (angle 90)	1.6555		Large Rodlike	0.6502
2nd order GLRLM	GLN (angle 90)	0.0725		Round and Regular	0.0942
	RLN (angle 90)	1.9178		Vascular	0.4437
	HGLRE (angle 135)	2.1866	Distribution	N/A	0.8289
	LGLRE (angle 135)	-1.1698		Clustered	0.4182
	RLN (angle 135)	-5.4927		Linear	1.3546
				Diffusely Scattered	0.6599

Table 6-4: Lasso selected features and their lambda coefficients of the mass subset

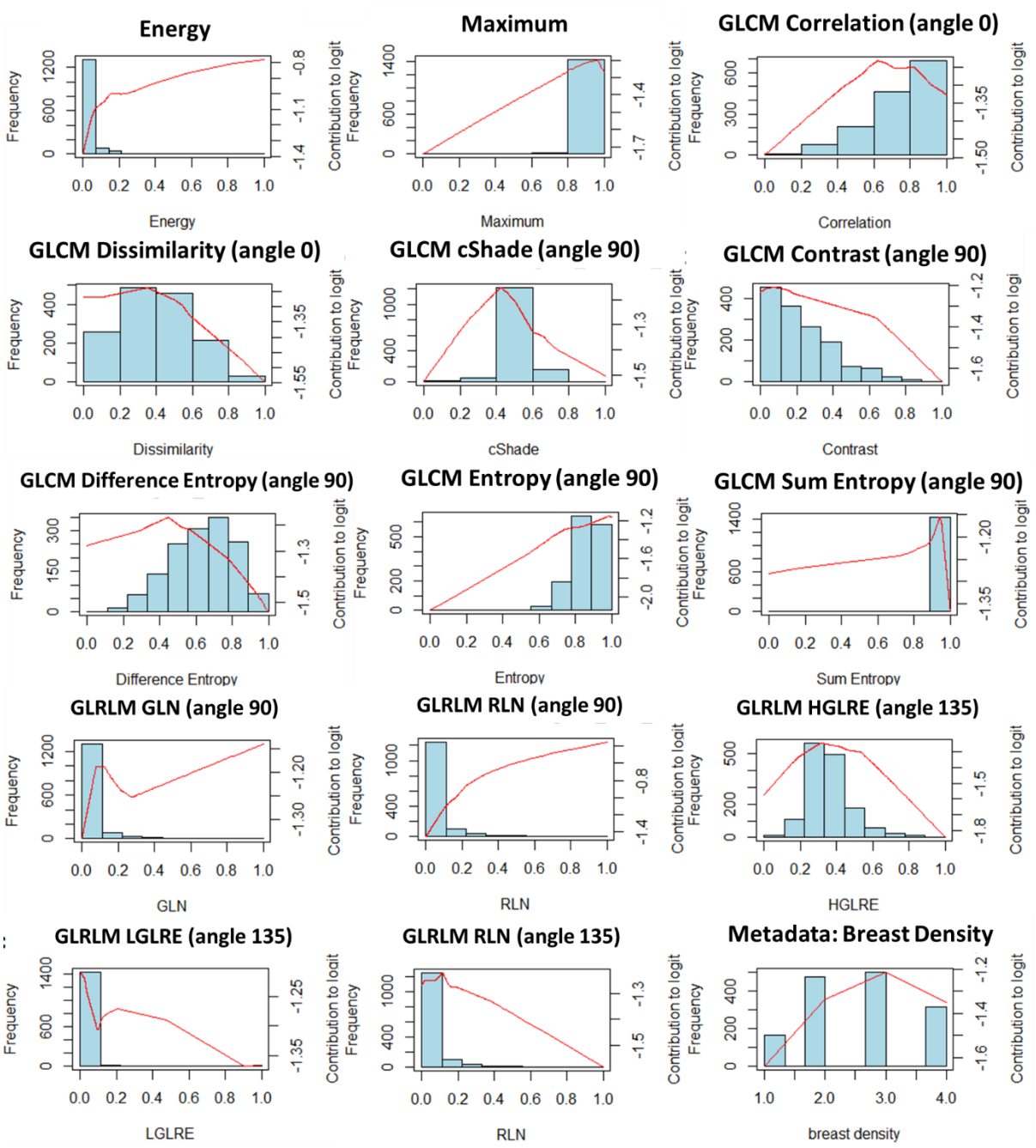


Figure 6-5: Partial responses for the continuous features (texture features, and breast density) from the calcification subset. For clarity, the values of the texture features have been scaled to be between 0 and 1.

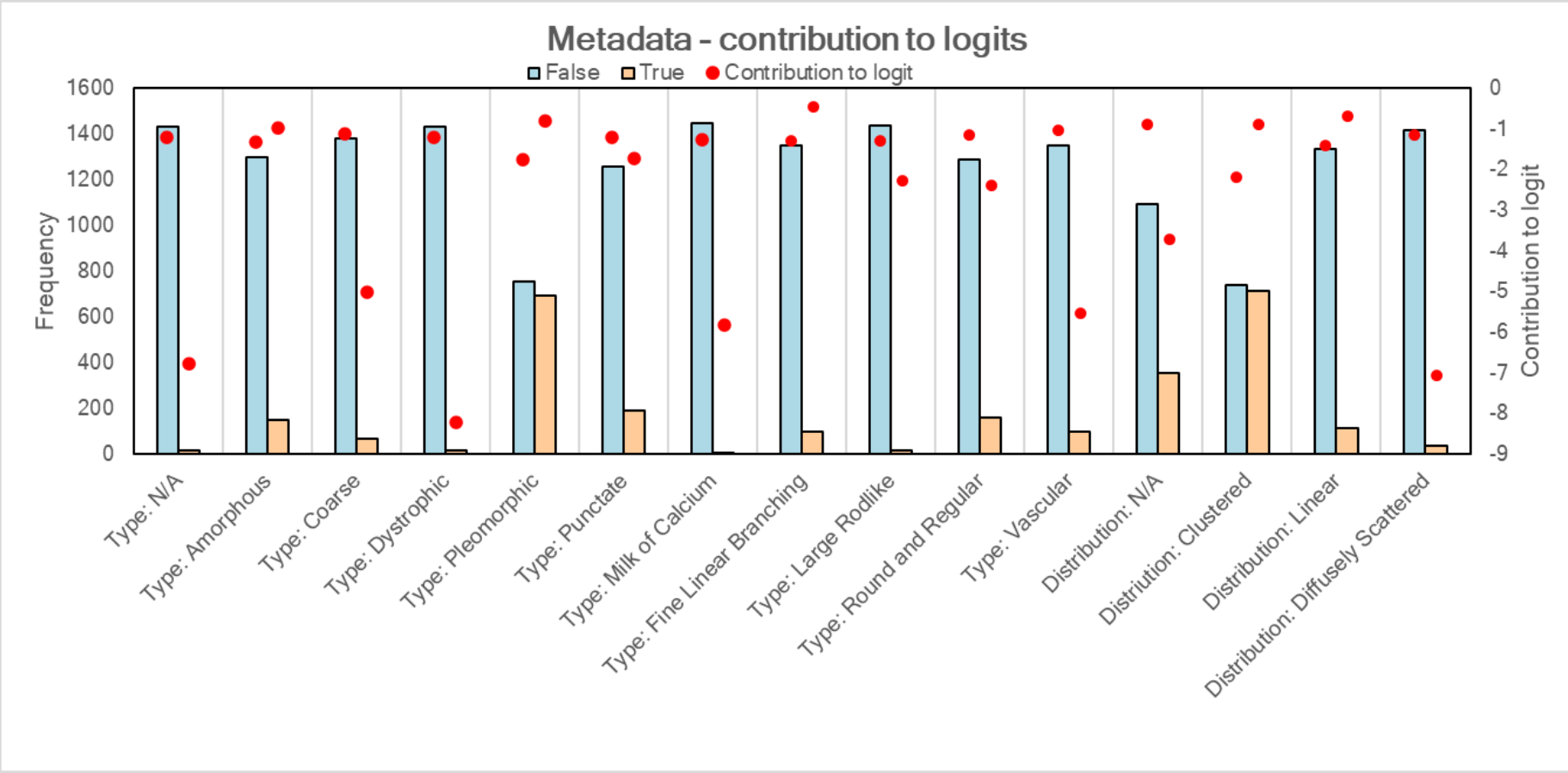


Figure 6-6: Partial responses (red points against y-axis on the right) for the categorical metadata for the calcification subset. Note, "True" denotes the given metadata aspect is true (=1).

6.3.4 Examples of cases with mammogram images

Examples of a case from each studied subset – one calcification and one mass – with their images and associated data are shown here, to focus on that case given data and their data.

6.3.4.1 Mass example

Figure 6-7 shows a mammogram containing a **malignant** mass. Some associated data is listed:

- Metadata:
    - Breast density: 3
    - Mass shape: Irregular
    - Mass margins: Ill-Defined and Spiculated
- Texture features (scaled 0-1):
    - Energy: 0.053
    - GLCM Inverse Variance (angle 135): 0.014
    - GLRLM SRHGLE (angle 45): 0.015

The breast density was not selected through the Lasso process but is included for completeness. The irregular shape was selected and follows as expected, as the contribution to the logit here tends to the malignant class. This also follows for the ill-defined margins and the spiculated margins, the latter much more severely against benign classifications. For all texture features, the story does not follow as clearly, however, the contribution to logit scale across the values is smaller than what it is for the metadata. This may suggest that although there is some impact (due to the Lasso selection), the metadata is performing more of the work in this case.

Figure 6-7: Mass example – mammogram of patient 15, left breast, MLO view. Malignant lesion.

6.3.4.2 Calcification example

Figure 6-8 shows a mammogram containing a **benign** calcification. Some associated data is listed:

- Metadata:
    - Breast density: 4
    - Calcification type: Pleomorphic
    - Calcification distribution: Linear
- Texture features (scaled 0-1):
    - Energy: 0.040
    - GLCM Correlation (angle 0): 0.002
    - GLRLM HGLRE (angle 135): 0.002

All variables noted here were selected by the Lasso process. The breast density category of 4, the highest, follows with the fact the lesion is benign. The

contribution leans more towards benign where the breast density category increases from 3 to 4. For both the pleomorphic calcification type and the linear calcification distribution, the difference between the contributions against the benign and malignant classes are negligible and so may not have as much of an impact for this case. All texture features noted follow in this case.

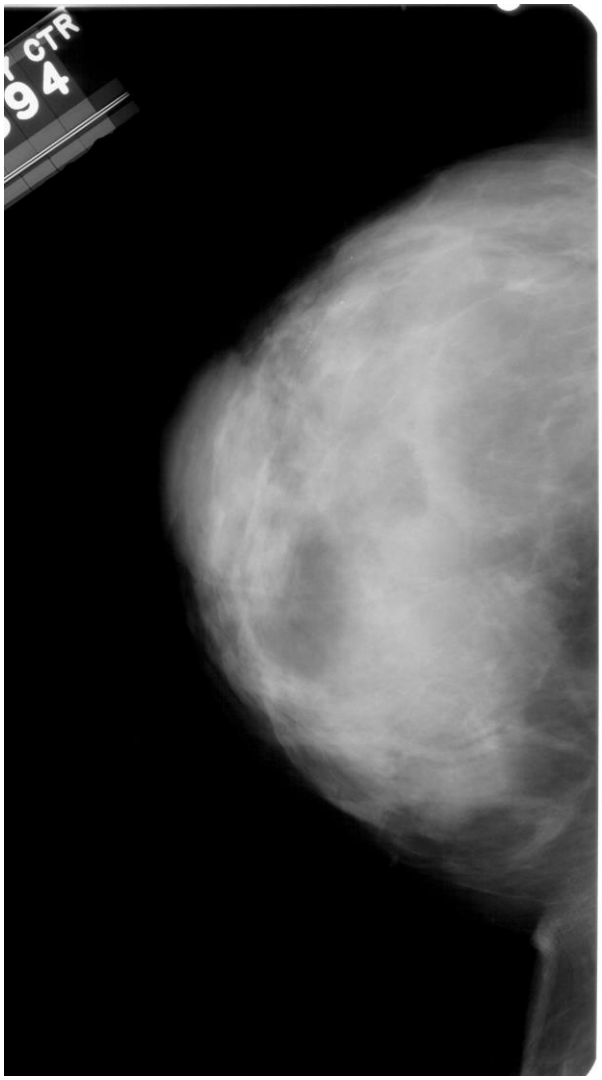


Figure 6-8: Calcification example – mammogram of patient 7, left breast, CC view. Benign lesion.

6.4 Conclusions

The aim of this work was to produce a more interpretable model showing the risk of malignancy of breast masses, with metadata and texture features. In this chapter, an MLP model with Lasso feature selection has provided more interpretable results and the ability to analyse important variables compared to the use of a standalone machine learning classifier. This result shows that a more

compact model can explain the relationship of each variable to the outcome to assist in clinical decision making, allowing for reasoning to take place alongside classification.

On its own, an MLP can identify non-linear maps, however, its interpretability can become a challenge. Explaining the contribution of a given variable in a dataset would traditionally be difficult, is important to clinicians. In this chapter, by letting one variable change at a time and keeping the others constant, MLP partial responses give rise to how each variable affects the outcome of the model predictions.

It is intended that this application can link machine learning capabilities with clinical reasoning, to enhance clinical decision-making. Partial responses can add insight to an AI-centred approach to cancer detection and classification.

7 Multimodal fusion including the use of CNN models to improve the classification of a breast cancer classifier

7.1 Introduction

This chapter studies methods to identify how best to combine available clinical data across different modalities to improve classification performance. Through building different types of multimodal fusion models, the aim is to identify suitable data fusion methods that can contextualise the output of a model, using machine learning techniques alongside important lesion metadata. These concepts have been applied throughout medical applications in the literature, including in pulmonary embolism [80], Alzheimer's [141], dermatology [142] and breast cancer [143].

This work can also be defined as "deep multimodal learning" where CNN features have been used within the data combinations. This allows for these traditionally higher-performing methods to be included within the analysis. Shared representations can be learned from the data throughout the different modalities and are scalable for the varied use of different modalities. Furthermore, much work in deep multimodal learning can review manually selected features, whereas this work starts with all available features, reducing through a feature selection process.

This work attempts to show how machine learning techniques can assist in a clinical setting. Through selecting the most useful features and performing the classification of lesions of malignancy – a typical clinical problem – it is intended that this can be seen as a further decision-making consideration for a multidisciplinary healthcare team.

This work looks at combining three sets of data that hold different aspects of information for the same cases within the CBIS-DDSM dataset: features from a CNN model, statistical texture features and associated clinical metadata about a given lesion.

7.2 Materials and methods

7.2.1 Data utilised from CBIS-DDSM

Due to their visual differences, the work is conducted separately on breast calcifications and breast masses with results presented for both lesion types. Three data subsets are used. Standalone classifiers are created on the CNN features [100], statistical texture features and the associated metadata with their results assessed. These are described in Chapter 3. Combinations of these, both two and three subsets, are also analysed. Table 7-1 shows the number of features available in each data subset being analysed, for each breast lesion type.

Data subset	Masses	Calcifications
CNN features	2048	
Statistical texture features	130	
Metadata	16	21

Table 7-1: Number of features available for each data subset, for each type of breast lesion.

7.2.2 Feature selection and classification

Feature Selection. As there are many features when including multiple data subsets, to meet the aim of enhancing the understanding of the classification process, a Lasso feature selection process is implemented in this data fusion approach. The intention here is to reduce prediction error compared to using all features, as well as indicate the more important variables in the process. A description of the Lasso is provided in Chapter 6. The results presented all use the Lasso parameter lambda which is within one standard error from the minimum. As described in the previous chapter, the lasso process looks to choose the simpler model with respect to error.

Classification. In this work, the logistic regression classifier has been used to discriminate between benign and malignant lesion cases with a probabilistic output. It uses the form shown in Equation 7-1.

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

Equation 7-1

7.2.3 Fusion methods

In this work, multimodal fusion will take pixel data analysis alongside other data types. It is intended that the combination of both the feature selection and the classification processes will meet the aim of contextualising the classification output of the process.

Combining data from different modalities into a single feature vector before training a machine learning classifier is known as “early fusion” or “data-level fusion” [144]. Huang et al. [21] refer to early fusion models as the process of joining multiple input modalities into a single input vector before feeding into one single machine learning model for training. In this work, the input modalities are concatenated into a single vector as appropriate for the data combination.

Late fusion, or decision-level fusion, is where decisions or probabilities from different classifiers are aggregated. Here, different modalities are trained separately, and the final decision is made through some form of aggregation. In this work, the mean average of the probabilities of the given classifiers provides the final prediction.

7.2.4 Experiments conducted

Various experiments have been conducted as comparison points to build on the standalone models by introducing forms of data fusion. There are three modalities of data being used throughout the process: CNN features, statistical texture features and metadata. The outcome measures are reported using the AUC measure.

Standalone experiments – feature selection and classification using only one set of data. These are shown in Figure 7-1.

Early Fusion A – concatenate **all** features depending on the combination being studied and *then* perform the feature selection and classification process. An example of this (for all three modalities) is shown in Figure 7-2 (a).

Early Fusion B – take the **selected** features from the "Standalone experiments," concatenate these and then perform the classification process (no further feature selection). An example of this (for all three modalities) is shown in Figure 7-2 (b).

Late Fusion – Take the predicted probabilities from the "Standalone experiments," and take the mean average. An example of this (for all three modalities) is shown in Figure 7-3.

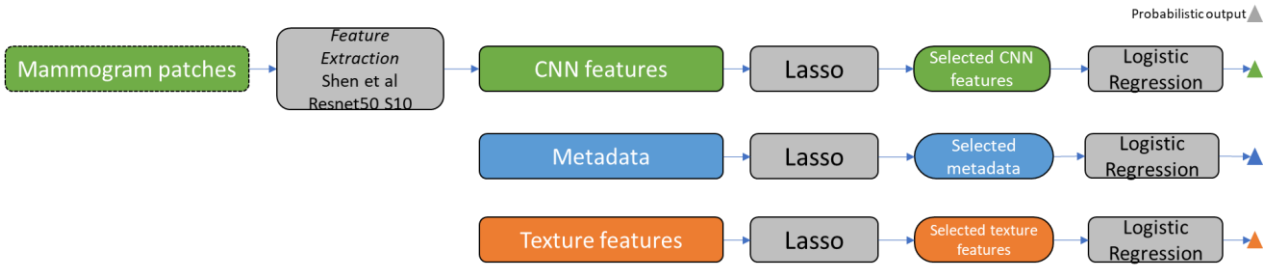


Figure 7-1: Standalone fusion examples. The design scheme follows in the other figures in this chapter.

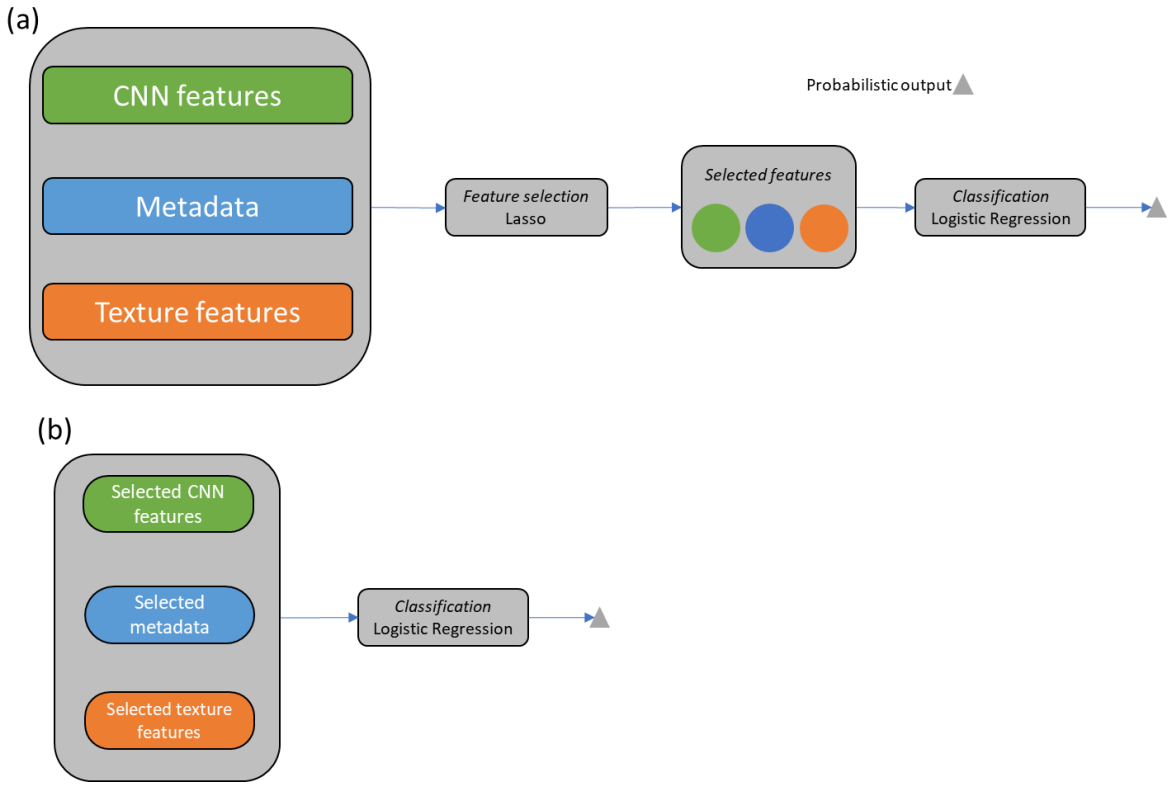


Figure 7-2: Early Fusion examples for all three modalities, where (a) is Early Fusion A and (b) is Early Fusion B. Similar for pairs of modalities.

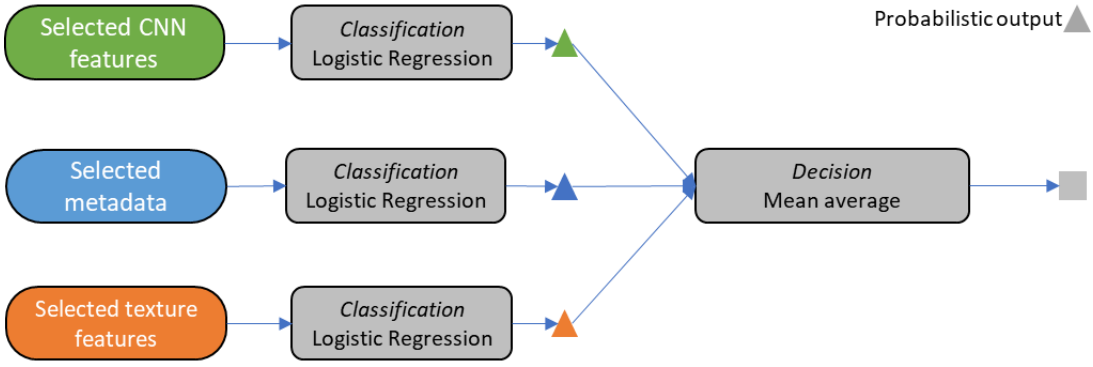


Figure 7-3: Late fusion example for all three modalities. Similar for pairs of modalities.

7.3 Results

As noted earlier, the results presented here have used the lambda argument of the Lasso that produces the model with the minimum number of variables within 1 standard error of lambda.

7.3.1 Standalone results

Table 7-2 shows the results for the standalone classifiers. The number and proportion of features selected through the Lasso process for each data subset are noted.

For both the masses and the calcifications the metadata alone attained the strongest performance with test set AUCs of 0.94 and 0.842 for the masses and calcifications, respectively. Alone, the statistical texture features performed the poorest with test set AUCs of 0.681 and 0.753 for masses and calcifications, respectively.

	Masses			Calcifications		
Data subset	**Number of selected features**	**Train AUC**	**Test AUC**	**Number of selected features**	**Train AUC**	**Test AUC**
CNN features only	668 (32.6%)	1	0.861	666 (32.5%)	0.999	0.808
Statistical texture features only	17 (13.1%)	0.711	0.681	21 (16.2%)	0.756	0.753
Metadata only	6 (37.5%)	0.866	0.94	9 (42.9%)	0.826	0.842

Table 7-2: Standalone classifier results.

7.3.2 Early Fusion A

Table 7-3 shows the results for the Early Fusion A process, where all features are concatenated, features are selected from this using the Lasso and these subsets of

features are used in the logistic regression classification process. The number and proportion of features selected through the Lasso process for each data subset are noted.

For both the masses and the calcifications the fusion of statistical texture features and metadata attained the strongest performance with test set AUCs of 0.947 and 0.888 for the masses and calcifications, respectively. The poorest performance is seen for the fusion of CNN features and statistical texture features with test set AUCs of 0.862 and 0.82 for masses and calcifications, respectively.

	Masses			Calcifications		
Data Combination	**Number of selected features**	**Train AUC**	**Test AUC**	**Number of selected features**	**Train AUC**	**Test AUC**
CNN features and metadata	551 (26.7%)	1	0.903	660 (32.2%)	1	0.844
CNN features and Statistical texture features	648 (29.8%)	1	0.862	626 (28.7%)	1	0.82
Statistical texture features and metadata	10 (6.8%)	0.884	0.947	25 (16.6%)	0.882	0.888
All three	564 (25.7%)	1	0.908	633 (28.8%)	1	0.821

Table 7-3: Early Fusion A results.

7.3.3 Overlapping features between standalone and Early Fusion A experiments

7.3.3.1 Masses

Table 7-4 shows the overlapping features between the standalone classifiers and the Early Fusion A classifiers for the mass lesions, as the latter are classifiers of combinations of data from the standalone. Several CNN features overlap between the standalone and fusion classifiers. These are the most difficult to interpret.

Statistical texture features can add some interpretability. In the best performing classifier for Early Fusion A for the masses, the statistical texture features and metadata classifier, maximum and GLCM Autocorrelation were selected by the Lasso. Autocorrelation measures the coarseness of an image, where higher values show greater concentration of low grey values within the ROI image.

In the CNN features and statistical texture feature classifier, both minimum and GLRLM RLN – Run Length Nonuniformity – were selected. The RLN feature

measures the similarity of run lengths throughout the image (where a run length is the length in which a number of consecutive pixels have the same grey level value).

For the classifier that uses all data subsets, minimum, maximum and GLCM Inverse Variance were selected. The GLCM Inverse Variance describes the inverse variance of the GLCM matrix, which examines the spatial relationship among pixels, defining how frequently a combination of pixels are present in an image. These variables, through the Lasso feature selection process, are seen as important in discriminating between a benign or a malignant lesion. An example of how this could be studied has been reviewed in the previous chapter on partial responses.

The metadata can provide the most interpretability out of the three data subsets. The same metadata features are selected suggesting they are important in discriminating between benign and malignant masses. To put the selected metadata features into clinical context, details from The Abnormal Mammogram [92] suggest that *irregular* masses (shape) indicates greater likelihood of malignancy, whereas the likelihood of malignancy with a *circumscribed* mass (margins) is lower but further work-up may be needed to verify the margins are completely circumscribed. Furthermore, *spiculated* margins may be more likely to be malignant.

7.3.3.2 Calcifications

Table 7-5 shows the Lasso selected features between the standalone classifiers and the Early Fusion A classifiers for the mass lesions.

More texture features were selected by the Lasso for the calcifications compared to the masses. In the best performing classifier, the metadata and statistical texture features classifier, entropy, minimum, GLRLM RLN and GLRLM SRLGLE – Short Run Low Grey Level Emphasis – were selected. Entropy measures the randomness within the image. GLRLM SRLGLE measures the joint distribution of shorter run lengths with lower grey-level values.

The CNN features and statistical texture features classifier selected energy, minimum, GLRLM RLN and GLRLM SRHGLE. The energy denotes the total magnitude of pixel values in a given image.

Where all three data subsets were used, the minimum, GLRLM RLN and GLRLM SRHGLE texture features were selected by the Lasso.

Regarding the results for the calcifications, like the masses in all cases where the metadata is analysed, the same metadata features are selected suggesting they are important in discriminating between benign and malignant calcifications. To put overlapping features from the calcification lesion data into clinical context, it is known [92] that *amorphous* calcifications are of immediate concern. Higher probability of malignancy includes *pleomorphic*, *linear,* and *fine linear branching*. However, many calcifications are so typical of a benign lesion that additional work is unnecessary.

Masses	*Early Fusion A*				
Standalone	**Overlapping features**	**Metadata and Statistical texture features**	**CNN features and Metadata**	**CNN features and Statistical texture features**	**All three**
	CNN features		(393 overlapped)	(580 overlapped)	(396 overlapped)
	Statistical texture features	1st order: Maximum 2nd order GLCM: Auto Correlation (angle 45)		1st order: Minimum 2nd order GLRLM: RLN (angle 45)	1st order: Maximum 1st order: Minimum 2nd order GLCM: Inverse Variance (angle 90)
	Metadata	Shape: Irregular Margins: Circumscribed Margins: Ill Defined Margins: Obscured Margins: Spiculated Margins: Microlobulated	Shape: Irregular Margins: Circumscribed Margins: Ill Defined Margins: Obscured Margins: Spiculated Margins: Microlobulated		Shape: Irregular Margins: Circumscribed Margins: Ill Defined Margins: Obscured Margins: Spiculated Margins: Microlobulated

Table 7-4: Overlapping features in the Mass lesion data, between the standalone classifiers and the Early Fusion A classifiers (all features concatenated, then feature selection). (GLCM: Grey Level Co-occurrence Matrix; GLRLM: Grey Level Run Length Matrix)

Calcifications	*Early Fusion A*				
Standalone	**Overlapping features**	**Metadata and Statistical texture features**	**CNN features and Metadata**	**CNN features and Statistical texture features**	**All three**
	CNN features		(524 CNN features overlapped)	(552 overlapped)	(489 overlapped)
	Statistical texture features	1st order: Entropy 1st order: Minimum 2nd order GLRLM: RLN (angle 90) 2nd order GLRLM: SRLGLE (angle 135)		1st order: Energy 1st order: Minimum 2nd order GLRLM: RLN (angle 90) 2nd order GLRLM: SRHGLE (angle 90)	1st order: Minimum 2nd order GLRLM: RLN (angle 90) 2nd order GLRLM: SRHGLE (angle 90)
	Metadata	Type: N/A Type: Amorphous Type: Pleomorphic Type: Fine Linear Branching Type: Round and Regular Distribution: N/A Distribution: Linear Distribution: Diffusely Scattered	Type: N/A Type: Amorphous Type: Pleomorphic Type: Fine Linear Branching Type: Round and Regular Distribution: N/A Distribution: Linear Distribution: Diffusely Scattered		Type: N/A Type: Amorphous Type: Pleomorphic Type: Fine Linear Branching Type: Round and Regular Distribution: N/A Distribution: Linear Distribution: Diffusely Scattered

Table 7-5: Overlapping features in the Calcification lesion data, between the standalone classifiers and the Early Fusion A classifiers (all features concatenated, then feature selection)

7.3.4 Early Fusion B

Table 7-6 shows the results for the Early Fusion B process, where the selected features from the standalone process are concatenated and classified with logistic regression (no further feature selection takes place).

For both the masses and the calcifications the fusion of statistical texture features & metadata attained the strongest performance with test set AUCs of 0.927 and 0.862 for the masses and calcifications, respectively. The poorest performance is seen for the fusion of CNN features & statistical texture features with test set AUCs of 0.828 and 0.744 for masses and calcifications, respectively.

	Masses			**Calcifications**		
Data Combination	**Number of selected features**	**Train AUC**	**Test AUC**	**Number of selected features**	**Train AUC**	**Test AUC**
CNN features and metadata	674	1	0.856	675	1	0.746
CNN features and Statistical texture features	685	1	0.828	687	1	0.744
Statistical texture features and metadata	23	0.9	0.927	30	0.893	0.862
All three	691	1	0.861	696	1	0.758

Table 7-6: Early Fusion B results.

7.3.5 Late Fusion

Table 7-7 shows the results for the late fusion process where the predicted probabilities from the standalone processes are combined and their mean average taken, leading to the reported AUC measure (no further feature selection takes place).

For both the masses and the calcifications the fusion of statistical texture features and metadata attained the strongest performance with test set AUCs of 0.906 and 0.862 for the masses and calcifications, respectively. The poorest performance is seen for the fusion of all three data subsets with test set AUCs of 0.785 and 0.726 for masses and calcifications, respectively.

Data Combination	Masses			Calcifications		
	Number of selected features	Train AUC	Test AUC	Number of selected features	Train AUC	Test AUC
CNN features and metadata	674	0.999	0.804	675	0.996	0.75
CNN features and Statistical texture features	685	0.999	0.817	687	0.996	0.754
Statistical texture features and metadata	23	0.881	0.906	30	0.858	0.862
All three	691	0.989	0.785	696	0.976	0.726

Table 7-7: Late Fusion results.

7.4 Discussion

By studying the predictive capabilities of both the standalone classifiers and fused classifiers as shown, it is intended that the reasoning behind the classification, including the important features chosen from the Lasso process, can be revealed.

To this extent, by including CNN features in the analysis it is possible to capture the inarguable strength of deep learning algorithms as noted throughout the literature. Using hand-crafted statistical texture features aligns with general aims in the field. The use of associated metadata furthers the interpretability aspect of the process including a view for the classifier to work “hand-in-hand” with clinicians. By considering knowledge from clinicians the predictive capabilities can be aided further.

The results of the standalone classifiers, detailed in Table 7-2 indicate that the approach taken in this chapter - Lasso feature selection process and logistic regression classifier – are reasonable. Although other classifiers could be used in place of logistic regression, in healthcare the interpretability of a machine learning algorithm is important [145]. The absence of explanation of the prediction within the decision-making process with “black box” machine learning models is somewhat of an issue, with interpretable machine learning models being seen as more useful instead [59].

On their own in the standalone classifiers, the statistical texture features are not enough to cause great impact. The CNN features attain more suitable results

however suffer with interpretability issues. Where focussing on the metadata alone, these attain the best results. However, these require work to attain the information whereas the CNN features and statistical texture features are attained from the images and do not. This is discussed further.

All results show that either standalone or fusion models can discriminate between benign and malignant lesions. In this work, the overlapping features between the standalone classifiers and the Early Fusion A classifiers have been highlighted. The CNN features are the hardest (near impossible) to interpret although the number overlapping is stated.

Through the feature selection process, models without CNN features – standalone and fusion models – are much smaller than if all features were used. For example, the best performing model, Early Fusion A – Statistical Texture Features and Metadata for the masses (Table 7-3) has a test AUC of 0.947 and is a model of 10 features. The next best performing uses all three data modalities and has a test AUC of 0.908 but with 564 features. The better performing model is arguably more interpretable without CNN features and much more lightweight without the need for images to be processed through a deep learning architecture in this case.

Early Fusion A models (Table 7-3) attain the best fusion results. This is likely because the Lasso feature selection process for each data combination has been able to assess and appropriately penalise features that do not contribute effectively or encounter inappropriate error. In particular, the statistical texture features and metadata combination

Early Fusion B, where the selected features from the standalone classifiers are concatenated and classified, does not perform as well as Early Fusion A, where all variables from each data subset are combined and then both the Lasso feature selection and the classification processes are performed. This is likely because in the Early Fusion B process, each standalone model has already had the most informative variables selected for each model alone. Whereas, with the Early Fusion A process, each variable can be reviewed against the outcome variable – pathology – providing more information at once.

The late fusion models provide some improvements on top of the standalone models, in some cases. For example, there is an improvement on the best result for the calcifications in Late Fusion (Statistical Texture Features and Metadata, test AUC 0.862) compared to standalone (Metadata only, test AUC 0.842), but this is not reflected in the masses. They do not perform better than the other fusion methods. This could be due to discrepancies in the predictions. The aggregate function was the mean average function, and it may be the case that it has been sensitive to differences in classification in some cases. This is not a strong issue, as the outcome measures are relatively strong, however, it may show a weakness in this method.

7.4.1 Why is this useful in a clinical setting?

The aim of this chapter has been for the multimodality data fusion methods to contextualise the classification process. In a clinical setting, it will be difficult to argue this using CNN features as the amount of mathematical processing in the methodology to get from input data to these features, is extensive. Although, their power as shown in the machine learning literature is included in this analysis. It is possible to build on this with the other data modalities studied, firstly with statistical texture features and then, most importantly, with the metadata.

The statistical texture features provide some interpretability. For example, through knowing which statistical texture features in the Early Fusion A classifiers are selected, it is possible to consider why. Consider a calcification, shown as white spots in a mammogram. Where the energy texture feature is selected, this could be higher where there are many calcification spots within an image. Further, GLRLM features may be affected due to an interruption in run lengths throughout an image, with run lengths being broken by the calcifications.

Consider a mass breast lesion. Rather than milk spots in an image, this will appear as a manifested lump within a mammogram. More severe breast masses – such as those with spiculated margins – are more likely to be malignant. Due to their vastly different margins which are not round but can present as “spiked”, neighbouring pixel values and lengths of similar grey levels are likely to be interrupted. In this case, GLCM and GLRLM values are likely to be significantly different than say, circumscribed margins where the shape is more likely to be round and smoother.

This can provide some insight into why the features are highlighted by the Lasso variable selector and insight into the classification process.

Interestingly, the selected metadata features are the same for each Early Fusion A classifier, for both the masses (Table 7-4) and the calcifications (Table 7-5) respectively. It is important to note at this stage that, while CNN features and statistical texture features are calculated directly from the images (in a sense, 'automatic'), the assigning of metadata is a manual process from a clinician. It could be suggested that these overlapping features are very important to the classifier. The Lasso feature selection process has effectively seen these as the features as the subset of features – or part of the subset of features – that can discriminate between benign and malignant classification, while minimising residual error.

It is acknowledged that the metadata is not provided until the images have been reviewed. In a real-world setting, the CNN features and statistical texture features could be calculated automatically, and the classification would take place. However, to improve this the metadata could be added retrospectively and the classification process for, say, statistical texture features and metadata (Early Fusion A) could take place instead, where the testing AUCs are 0.947 and 0.888 for masses and calcifications, respectively. Adding this associated metadata can support clinical thinking and reasoning. It can be argued that the machine is working in cooperation with clinicians, through appending new knowledge.

With the inclusion of the associated metadata in the analysis, strong improvement to the predictive capability can be seen. This can be seen as cooperation between the machine learning algorithm and clinicians. By including knowledge from a clinical setting, we can further aid the "thinking of the classifier" and exploit clinical impact. Consider the best models – both Early Fusion A for masses and calcifications, using statistical texture features and metadata. As described these classifiers provide the most interpretability as they do not include the CNN features. Where the statistical texture features can provide some interpretability, appending the metadata furthers this. Here, the machine can work "hand in hand" with the clinician.

It is reported that between radiologists, observations made on mammograms can differ substantially [6]. Methods used in this work can attempt to assist with this. Leveraging data fusion techniques has improved performance and where appropriate, the ability to put the prediction into context.

The complexity of the lesion metadata was limited to what a clinician may assign and is not automatically detected. It does not include other measures, say blood tests for a patient; the furthest “patient-level” data available is the breast density. Within the literature, there is work within multimodality fusion that looks at patient electronic health records [80], [143] which can include age, family history and lifestyle factors that are seen to be factors in breast cancer causes [90]. Future work could benefit from using such large volumes of data in a similar fashion as presented here, including a feature selection process.

7.5 Conclusions

In this chapter, a framework using multimodality data fusion for breast cancer classification has been presented. This work studied techniques that utilised features attained from a deep learning model, hand-crafted statistical texture features and lesion-level metadata.

The aim of this work was to identify suitable data fusion models that holds reasonable predictive capabilities and the ability to contextualise the decision-making process. The Lasso feature selection process followed by logistic regression classification upholds interpretability and avoids “black box” classification, and studying the overlapping features ensures important features are recognised for their usefulness.

Part of this work, particularly classifiers that include the use of associated lesion metadata, look to reflect clinical thinking. Multimodal data fusion models in this work attempt to exploit the amount of available information for a given case and show how building on imaging data with clinical knowledge can further aid the classification process. This could be seen as an extra tool in the arsenal of a clinician in the setting and assist with decision making.

Future work could consider the use of patients’ electronic health records such as patient information and test results which could introduce further benefits to diagnosis and triage.

8 A voting ensemble method to assist the diagnosis of prostate cancer using multiparametric MRI

Embargoed

Embargoed

Embargoed

Embargoed

Embargoed

Embargoed

Embargoed

Embargoed

Embargoed

Embargoed

Embargoed

9 Conclusions

9.1 Review and conclusions of the work presented within the thesis

This thesis has presented a set of tools using machine learning techniques that attempt to bridge the gap between the clinician and the machine, in the application of cancer classification. Much of this work can be seen to improve the diagnostic arsenal available to multidisciplinary healthcare teams, including possible moves towards a more automated triage system where a method could be deemed effective. Three categories of methods and applications have been presented:

- The first section defined the Fisher Information metric methodology, leading to a mathematically robust representation of the nearest neighbour structure of clinical data. This investigated the use of a distance-based metric which relied upon the distances between observations. This led to the creation of a latent data space in which every patient case could be visualised, leading to a "patient like me" approach for the addition of new data unseen to the classifier. Using this allows for a more evidence-based approach to triaging new cases.
- The second section studied the effect of different variables in a clinical dataset and how they impacted the outcome of the predictive model. This analysis of an MLP model is an attempt to provide understanding and "open the black box" by reviewing the extent to which different values of given variables affect the contribution the logit.
- The third section reviewed multimodality data fusion to improve and contextualise the predictive capability of a classifier by adding further information to the classifier in different ways. This work attempted to exploit the abundance of available clinical data to aid the decision-making process.

The first section (Chapters 4 and 5) presented a robust mathematical representation of clinical data under two applications – using statistical texture features and using features extracted through deep learning. This was to consider how decisions are made by multidisciplinary healthcare teams using various

techniques and evidence. For this, a defined metric was created to “visualise the thinking” of an MLP classifier by projecting trained cases into a latent space. In effect, this method takes what was inherently “learned” by the weights and biases of the classifier. This led to a new “patient like me” approach for new patient cases. Where a machine learning model can assess new cases, the FIN methodology has been extended in this thesis to visualise new cases within the embedding with multidimensional scaling and similarity distance calculations.

Standalone MLP or CNN classifiers can provide probabilistic output albeit with limited explanations as to why. It is important to provide personalised care in cancer diagnosis and treatment and this work has presented a machine learning, evidence-based solution to this important and sensitive area. The Fisher Information Metric “makes distances a reliable measure of how similar samples really are with regard to the underlying posterior class probabilities.” [30]

This work found that it is possible to conduct a “patient like me” approach, from training an MLP to discriminate between benign and malignant tumours, calculate a metric which defines a Riemannian space where the distances between each point are a measure of similarity, calculate the magnitude of similarity between each point and use multidimensional scaling to project this into a visualisation. Then, test cases were projected and a demonstration of how this could be used – including with patient metadata to ascertain further clinical insight – can be a machine learning application to a sensitive and important clinical problem.

As a diagnostic tool, the Fisher Information Network visualisations can be seen to be both independent from clinicians yet able to assist. Where a clinical team will have their own knowledge of radiology and anatomy, the machine learning models have also ascertained their own knowledge through the model training processes as well as being able to quantify and assess their performance.

The next section (Chapter 6) studied the partial responses of a neural network classifier. This was proposed to assess the changes in values of variables in a clinical dataset, to evaluate the changes within variables that affect the predictive outcome. This work used a methodology that involves an MLP classifier and Lasso

feature selector to study important variables and how changes in values can affect the contribution to the logit of an MLP, affecting the final classification.

While an MLP on its own can identify non-linear mappings, its interpretability can become a challenge. This work has attempted to explain the contributions of given variables throughout the dataset, which is an important clinical aspect. Knowing exactly what changes can lead to a different outcome can be incredibly useful in a healthcare setting with wide-reaching benefits. In this study, by letting one variable change at a time and keeping the rest constant, the partial responses of the MLP give rise to how each variable affects the outcome of the models' predictive capabilities. This work can link machine learning capabilities with clinical reasoning, to enhance the decision-making process.

The final section (Chapters 7 and 8) reviewed the use of multimodality data fusion in clinical settings to contextualise the decision-making process and use the abundance of available data as a starting point to identify a suitable grouping of data to do this effectively. Here a combination of three data subsets on the same data was used with a logistic regression classifier, a traditionally interpretable and preferred clinical model to assess this hypothesis.

This work has shown that fusing different sets of data that relate to the same case and using feature selection to utilise the most important variables, suitable data combinations can exploit the information provided to the models to enhance predictions from just standalone models. In a bid to reflect clinical thinking, the best model in this work can be seen to augment automatically calculated features – statistical texture features – and clinically informed input – associated lesion metadata. This could be further augmented in a different application with electronic health records and patient-level metadata. This work look towards clinicians and machines working hand-in-hand with the best knowledge from both developing further insight and appending new knowledge for the task.

9.2 Future work statement

This research leads to further avenues for future work, a possible example in active machine learning is discussed.

9.2.1 Active learning

It is known that clinicians may not agree with other clinicians when diagnosing tumours [6], alongside other human-based issues including fatigue and workload [157]. In any case, strongly performing models require humans to train the model. Throughout all this work, the labels provided within the dataset have been treated as read and correct.

Furthermore, at the time of presentation, the patient will not have a classification from a biopsy of benign or malignant as this is invasive – the methods proposed here use only mammogram screening images. Where necessary after the screening, a biopsy is taken. At this stage, a patient's data would be completed in the context of the work presented in the thesis. For the FIN map, say, these new and updated results could be added to the trained embedding.

A limitation of the FIN methodology work is that the calculation of the pairwise distances is hugely expensive in its computation. The distance of each point from each other point is calculated. The use of active learning here to further improve the embedding could append strong clinical impact to, say, national screening programmes [158]. Scalable implementations of the Fisher methodology require either powerful or Cloud Computing applications [101]. Developments to this could provide a more automated triage system where the predictive capabilities of the models were deemed to be strong enough.

Appendix 1 – Texture features selected in Partial Responses chapter

First order features These features describe the grey level distribution of the image.		
Name	*Equation*	*Description*
Energy	$E = \sum_{x=1}^{X}\sum_{y=1}^{Y}\sum_{z=1}^{Z} I(x,y,z)^2$	
Kurtosis	$Y^2 = \frac{1}{XYZ}\sum_{x=1}^{X}\sum_{y=1}^{Y}\sum_{z=1}^{Z}\left\{\left[\frac{I(x,y,z)-\mu}{\sigma}\right]\right\}^4 - 3$	Kurtosis denotes the sharpness of the histogram.
Maximum	$I_{max} = \max\{I(x,y,z)\}$	
Minimum	$I_{min} = \min\{I(x,y,z)\}$	

Second order GLCM features (Grey-Level Co-Occurrence matrix based features)

Let:

$P(i,j)$ be the co-occurence matrix.

N_g be the number of discrete intensity levels in the image.

μ be the mean of $P(i,j)$

$\mu_x(i)$ be the mean of row i

$\mu_y(j)$ be the mean of column j

$\sigma_x(i)$ be the standard deviation of row i

$\sigma_y(j)$ be the standard deviation of column j

$$p_x(i) = \sum_{j=1}^{N_g} P(i,j)$$

$$p_y(i) = \sum_{i=1}^{N_g} P(i,j)$$

$$p_{x+y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} P(i,j), i+j = k, k = 2,3,\dots,2N_g$$

$$p_{x-y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} P(i,j), |i-j| = k, k = 0,1,\dots,N_g - 1$$

Name	*Equation*	*Description*

Entropy	$entropy = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} P(i,j)\log_2[P(i,j)]$	Measures randomness, lower values for smoother images. Homogenous images have high entropy and vice versa.
Contrast	$contrast = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \lvert i-j\rvert^2 P(i,j)$	Measures local intensity variation. High for images with high contrast.
Inverse Variance	$inverse\ variance = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{P(i,j)}{\lvert i-j\rvert^2}$, $i \neq j$	Variance puts relatively high weights on the elements that differ from the average value of $P(i,j)$.
Correlation	$correlation = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} ijP(i,j) - \mu_i(i)\mu_j(j)}{\alpha_x(i)\alpha_y(j)}$	Correlation denotes the grey level linear dependence between pixels, at the specified positions, relative to each other. (correlation between pixels in two different directions)
Dissimilarity	$dissimilarity = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \lvert i-j\rvert P(i,j)$	Measures the distance between pairs of pixels in an image.

Sum Entropy	$$sum\ entropy = -\sum_{i=2}^{2N_g} P_{x+y}(i)\log_2[P_{x+y}(i)]$$	Measures the disorder related to the grey level-sum distribution of an image.
Difference Entropy	$$difference\ entropy = \sum_{i=0}^{N_g-1} P_{x-y}(i)\log_2[P_{x-y}(i)]$$	Measures the disorder related to the grey level difference distribution of an image.
Cluster Shade	$$cluster\ shade = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}[i+j-\mu_x(i) - \mu_y(j)]^3 P(i,j)$$	Measure of skewness of GLCM matrix. Higher value means the image is asymmetric.
Auto correlation	$$auto\ correlation = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} ijP(i,j)$$	Auto correlation denotes the coarseness of an image, evaluating the linear spatial relationships between texture primitives.

Second order GLRLM features (Grey-Level Run-Length matrix based features)

$p(i,j|\theta)$ is the $(i,j)th$ entry in the given run-length matrix, p, for a direction θ.

N_g: the number of discrete intensity values in the image.

N_r: the number of different run lengths.

N_p: the number of voxels in the image.

Name	*Equation*	*Description*
Grey level non-uniformity	$GLN = \frac{\sum_{i=1}^{N_g}\left[\sum_{j=1}^{N_r} p(i,j\|\theta)\right]^2}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j\|\theta)}$	GLN denotes the similarity of grey level intensity values within the image. This value is low if the intensity values are alike.
High grey level run emphasis	$HGLRE = \frac{\sum_{i=1}^{N_g}\left[\sum_{j=1}^{N_r} i^2 p(i,j\|\theta)\right]^2}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j\|\theta)}$	HGLRE denotes the distribution of high grey level values. This value is high for the image with high grey level values.
Low grey level run emphasis	$LGLRE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\left[\frac{p(i,j\|\theta)}{i^2}\right]}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j\|\theta)}$	LGLRE measures the distribution of low grey level values. It is high for images with low grey level values.

Run length non uniformity	$RLN = \frac{\sum_{i=1}^{N_g}\left[\sum_{j=1}^{N_r} p(i,j\|\theta)\right]^2}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j\|\theta)}$	RLN measures the similarity of the length of the runs throughout the image. It is low if the run lengths are alike.
Short run high grey level emphasis	$SRHGLE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}[\frac{p(i,j\|\theta)i^2}{j^2}]}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j\|\theta)}$	SRHGLE measures the joint distribution of short runs and high grey level values. It is high for images with many short runs and high grey level values.

References

[1] World Health Organization, 'Breast cancer', *WHO website*, 2018.

[2] 'NHS England » Review of national cancer screening programmes in England'. https://www.england.nhs.uk/publication/terms-of-reference-review-national-cancer-screening-programmes-england/ (accessed Aug. 30, 2021).

[3] R. E. Bird, T. W. Wallace, and B. C. Yankaskas, 'Analysis of cancers missed at screening mammography.', *Radiology*, vol. 184, no. 3, pp. 613–617, Sep. 1992, doi: 10.1148/radiology.184.3.1509041.

[4] J. Wyatt and D. Spiegelhalter, 'Field trials of medical decision-aids: potential problems and solutions.', *Proc Annu Symp Comput Appl Med Care*, pp. 3–7, 1991.

[5] 'A comparison of cancer detection rates achieved by breast cancer screening programmes by number of readers, for one and two view mammography: results from the UK National Health Service breast screening programme', *J Med Screen*, vol. 5, no. 4, pp. 195–201, Dec. 1998, doi: 10.1136/jms.5.4.195.

[6] E. Molins, F. Macià, F. Ferrer, M.-T. Maristany, and X. Castells, 'Association between Radiologists' Experience and Accuracy in Interpreting Screening Mammograms', *BMC Health Services Research*, vol. 8, no. 1, p. 91, Apr. 2008, doi: 10.1186/1472-6963-8-91.

[7] S. Saadatmand *et al.*, 'MRI versus mammography for breast cancer screening in women with familial risk (FaMRIsc): a multicentre, randomised, controlled trial', *The Lancet Oncology*, vol. 20, no. 8, pp. 1136–1147, Aug. 2019, doi: 10.1016/S1470-2045(19)30275-X.

[8] 'Early detection of breast cancer by surveillance | Information for the public | Familial breast cancer: classification, care and managing breast cancer and related risks in people with a family history of breast cancer | Guidance | NICE'. https://www.nice.org.uk/guidance/cg164/ifp/chapter/early-detection-of-breast-cancer-by-surveillance (accessed Aug. 30, 2021).

[9] 'Variation in Mammographic Breast Density Assessments Among Radiologists in Clinical Practice: A Multicenter Observational Study: Annals of Internal Medicine: Vol 165, No 7'. https://www.acpjournals.org/doi/abs/10.7326/m15-2934 (accessed Aug. 15, 2021).

[10] M. H. Gail *et al.*, 'Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually', *JNCI: Journal of the National Cancer Institute*, vol. 81, no. 24, pp. 1879–1886, Dec. 1989, doi: 10.1093/jnci/81.24.1879.

[11] T. Anothaisintawee, Y. Teerawattananon, C. Wiratkapun, V. Kasamesup, and A. Thakkinstian, 'Risk prediction models of breast cancer: a systematic review of model performances', *Breast Cancer Res Treat*, vol. 133, no. 1, pp. 1–10, May 2012, doi: 10.1007/s10549-011-1853-z.

[12] W. Xu, W. Liu, L. Li, G. Shao, and J. Zhang, 'Identification of Masses and Microcalcifications in the Mammograms Based on Three Neural Networks: Comparison and Discussion', in *2008 2nd International Conference on Bioinformatics and Biomedical Engineering*, May 2008, pp. 2299–2302. doi: 10.1109/ICBBE.2008.907.

[13] C. Abirami, R. Harikumar, and S. R. S. Chakravarthy, 'Performance analysis and detection of micro calcification in digital mammograms using wavelet

features', in *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, Mar. 2016, pp. 2327–2331. doi: 10.1109/WiSPNET.2016.7566558.
[14] M. A. Mazurowski, J. Y. Lo, B. P. Harrawood, and G. D. Tourassi, 'Mutual information-based template matching scheme for detection of breast masses: From mammography to digital breast tomosynthesis', *Journal of Biomedical Informatics*, vol. 44, no. 5, pp. 815–823, Oct. 2011, doi: 10.1016/J.JBI.2011.04.008.
[15] M. Elter and E. Haßlmeyer, 'A knowledge-based approach to the CADx of mammographic masses', Mar. 2008, vol. 6915, p. 69150L. doi: 10.1117/12.770135.
[16] C.-H. Wei, Y. Li, and P. J. Huang, 'Mammogram retrieval through machine learning within BI-RADS standards', *Journal of Biomedical Informatics*, vol. 44, no. 4, pp. 607–614, Aug. 2011, doi: 10.1016/J.JBI.2011.01.012.
[17] A. Akselrod-Ballin *et al.*, 'Predicting Breast Cancer by Applying Deep Learning to Linked Health Records and Mammograms', *Radiology*, vol. 292, no. 2, pp. 331–342, Aug. 2019, doi: 10.1148/radiol.2019182622.
[18] C. D. Lehman *et al.*, 'Mammographic Breast Density Assessment Using Deep Learning: Clinical Implementation', *Radiology*, vol. 290, no. 1, pp. 52–58, Jan. 2019, doi: 10.1148/radiol.2018180694.
[19] D. A. Ragab, M. Sharkas, S. Marshall, and J. Ren, 'Breast cancer detection using deep convolutional neural networks and support vector machines.', *PeerJ*, vol. 7, p. e6201, 2019, doi: 10.7717/peerj.6201.
[20] M. Reyes *et al.*, 'On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities', *Radiology: Artificial Intelligence*, vol. 2, no. 3, p. e190043, May 2020, doi: 10.1148/ryai.2020190043.
[21] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, 'Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines', *npj Digital Medicine*, vol. 3, no. 1, Art. no. 1, Oct. 2020, doi: 10.1038/s41746-020-00341-z.
[22] 'Breast screening: programme overview', *GOV.UK*. https://www.gov.uk/guidance/breast-screening-programme-overview (accessed Aug. 30, 2021).
[23] 'Coronavirus: "More than two million" waiting for cancer care in UK', *BBC News*, Jun. 01, 2020. Accessed: Aug. 30, 2021. [Online]. Available: https://www.bbc.com/news/health-52876999
[24] S. W. Duffy *et al.*, 'Effect of mammographic screening from age 40 years on breast cancer mortality (UK Age trial): final results of a randomised, controlled trial', *The Lancet Oncology*, vol. 21, no. 9, pp. 1165–1172, Sep. 2020, doi: 10.1016/S1470-2045(20)30398-3.
[25] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, 'A curated mammography data set for use in computer-aided detection and diagnosis research', *Scientific Data*, vol. 4, no. 1, pp. 170–177, Dec. 2017, doi: 10.1038/sdata.2017.177.
[26] A. Bellet, A. Habrard, and M. Sebban, 'Metric Learning', *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 9, no. 1, pp. 1–151, Jan. 2015, doi: 10.2200/S00626ED1V01Y201501AIM030.
[27] C. M. Bishop, *Pattern recognition and machine learning*, 1st ed. Cambridge: Springer, 2006.

[28] S. Wold, K. Esbensen, and P. Geladi, ‘Principal component analysis’, *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1–3, pp. 37–52, Aug. 1987, doi: 10.1016/0169-7439(87)80084-9.
[29] K. Pearson, ‘LIII. On lines and planes of closest fit to systems of points in space’, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, Nov. 1901, doi: 10.1080/14786440109462720.
[30] H. Ruiz, ‘Fisher networks: A principled approach to retrieval-based classification’, doctoral, Liverpool John Moores University, 2013. doi: 10.24377/LJMU.t.00004371.
[31] T. F. Cox, *Multidimensional scaling*, 2nd ed. Boca Raton, Fla. ; London: Chapman & Hall/CRC, 2001.
[32] L. Van Der Maaten and G. Hinton, ‘Visualizing Data using t-SNE’, 2008.
[33] S. T. Roweis and L. K. Saul, ‘Nonlinear Dimensionality Reduction by Locally Linear Embedding’, *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000, doi: 10.1126/science.290.5500.2323.
[34] C. M. Bishop, M. Svensén, and C. K. I. Williams, ‘GTM: The Generative Topographic Mapping’, *Neural Computation*, vol. 10, no. 1, pp. 215–234, Jan. 1998, doi: 10.1162/089976698300017953.
[35] ‘Flexible Metric Nearest Neighbor Classification | Department of Statistics’. https://statistics.stanford.edu/research/flexible-metric-nearest-neighbor-classification (accessed Aug. 28, 2021).
[36] T. Cover and P. Hart, ‘Nearest neighbor pattern classification’, *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.
[37] K. Q. Weinberger and L. K. Saul, ‘Distance Metric Learning for Large Margin Nearest Neighbor Classification’, *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Jun. 2009.
[38] T. S. Jaakkola and D. Haussler, ‘Exploiting generative models in discriminative classifiers’, in *Proceedings of the 1998 conference on Advances in neural information processing systems II*, Cambridge, MA, USA, Jul. 1999, pp. 487–493.
[39] S. Amari and S. Wu, ‘Improving support vector machine classifiers by modifying kernel functions’, *Neural Networks*, vol. 12, no. 6, pp. 783–789, Jul. 1999, doi: 10.1016/S0893-6080(99)00032-5.
[40] S. Kaski and J. Sinkkonen, ‘Metrics that learn relevance’, in *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, Jul. 2000, vol. 5, pp. 547–552 vol.5. doi: 10.1109/IJCNN.2000.861526.
[41] S. Kaski, J. Sinkkonen, and J. Peltonen, ‘Bankruptcy analysis with self-organizing maps in learning metrics’, *IEEE Transactions on Neural Networks*, vol. 12, no. 4, pp. 936–947, Jul. 2001, doi: 10.1109/72.935102.
[42] A. Tsymbal, ‘HeC CaseReasoner: NGraphs and Learning Discriminative Distances’, Jan. 2009. [Online]. Available: https://indico.cern.ch/event/49944/attachments/962313/1366268/01_06_CaseReasoner.pdf
[43] A. Tsymbal, M. Huber, and S. K. Zhou, ‘Learning Discriminative Distance Functions for Case Retrieval and Decision Support’, *Transactions on Case-based Reasoning*, vol. 3, no. 1, pp. 1–16, Oct. 2010.

[44] G. T. Toussaint, ‘The relative neighbourhood graph of a finite planar set’, *Pattern Recognition*, vol. 12, no. 4, pp. 261–268, Jan. 1980, doi: 10.1016/0031-3203(80)90066-7.
[45] M. Girvan and M. E. J. Newman, ‘Community structure in social and biological networks’, *PNAS*, vol. 99, no. 12, pp. 7821–7826, Jun. 2002, doi: 10.1073/pnas.122653799.
[46] L. Breiman, ‘Random Forests’, *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[47] T. Hastie and R. Tibshirani, ‘Generalized Additive Models’, *Statistical Science*, vol. 1, no. 3, pp. 297–310, 1986.
[48] P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman, ‘Sparse additive models’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 71, no. 5, pp. 1009–1030, 2009, doi: 10.1111/j.1467-9868.2009.00718.x.
[49] J. Chhatwal *et al.*, ‘A Logistic Regression Model Based on the National Mammography Database Format to Aid Breast Cancer Diagnosis’, *AJR. American journal of roentgenology*, vol. 192, no. 4, p. 1117, Apr. 2009, doi: 10.2214/AJR.07.3345.
[50] I. F. Gareen and C. Gatsonis, ‘Primer on Multiple Regression Models for Diagnostic Imaging Research’, *Radiology*, vol. 229, no. 2, pp. 305–310, Nov. 2003, doi: 10.1148/radiol.2292030324.
[51] V. Naumovich. Vapnik and V. N., *The nature of statistical learning theory*. Springer, 1995. Accessed: Aug. 16, 2018. [Online]. Available: https://dl.acm.org/citation.cfm?id=211359
[52] V. V. Belle, B. V. Calster, S. V. Huffel, J. A. K. Suykens, and P. Lisboa, ‘Explaining Support Vector Machines: A Color Based Nomogram’, *PLOS ONE*, vol. 11, no. 10, p. e0164568, Oct. 2016, doi: 10.1371/journal.pone.0164568.
[53] Leo. Breiman, *Classification and regression trees*. Chapman & Hall, 1993. Accessed: May 12, 2019. [Online]. Available: https://books.google.co.uk/books/about/Classification_and_Regression_Trees.html?id=JwQx-WOmSyQC&redir_esc=y
[54] C. Olah, A. Mordvintsev, and L. Schubert, ‘Feature Visualization’, *Distill*, vol. 2, no. 11, p. e7, Nov. 2017, doi: 10.23915/distill.00007.
[55] K. Simonyan, A. Vedaldi, and A. Zisserman, ‘Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps’, *arXiv:1312.6034 [cs]*, Apr. 2014, Accessed: Aug. 29, 2021. [Online]. Available: http://arxiv.org/abs/1312.6034
[56] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, ‘Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization’, *Int J Comput Vis*, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.
[57] M. D. Zeiler and R. Fergus, ‘Visualizing and Understanding Convolutional Networks’, in *Computer Vision – ECCV 2014*, Cham, 2014, pp. 818–833. doi: 10.1007/978-3-319-10590-1_53.
[58] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, ‘SmoothGrad: removing noise by adding noise’, *arXiv:1706.03825 [cs, stat]*, Jun. 2017, Accessed: Aug. 29, 2021. [Online]. Available: http://arxiv.org/abs/1706.03825

[59] C. Rudin, ‘Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead’, *Nat Mach Intell*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.
[60] J. H. Friedman, ‘Greedy function approximation: A gradient boosting machine.’, *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/aos/1013203451.
[61] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, ‘Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation’, *Journal of Computational and Graphical Statistics*, vol. 24, no. 1, pp. 44–65, Jan. 2015, doi: 10.1080/10618600.2014.907095.
[62] M. T. Ribeiro, S. Singh, and C. Guestrin, ‘“Why Should I Trust You?”: Explaining the Predictions of Any Classifier’, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.
[63] J. Gao, P. Li, Z. Chen, and J. Zhang, ‘A Survey on Deep Learning for Multimodal Data Fusion’, *Neural Computation*, vol. 32, no. 5, pp. 829–864, May 2020, doi: 10.1162/neco_a_01273.
[64] K. Ali, ‘On the Link between Error Correlation and Error Reduction in Decision Tree Ensembles’, 1995.
[65] T. K. Ho, J. J. Hull, and S. N. Srihari, ‘Decision combination in multiple classifier systems’, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 1, pp. 66–75, Jan. 1994, doi: 10.1109/34.273716.
[66] K.-H. Thung, P.-T. Yap, and D. Shen, ‘Multi-stage Diagnosis of Alzheimer’s Disease with Incomplete Multimodal Data via Multi-task Deep Learning’, *Deep Learn Med Image Anal Multimodal Learn Clin Decis Support (2017)*, vol. 10553, pp. 160–168, Sep. 2017, doi: 10.1007/978-3-319-67558-9_19.
[67] G. An *et al.*, ‘Comparison of Machine-Learning Classification Models for Glaucoma Management’, *Journal of Healthcare Engineering*, vol. 2018, p. e6874765, Jun. 2018, doi: 10.1155/2018/6874765.
[68] L. Tong, Y. Sha, and M. D. Wang, ‘Improving Classification of Breast Cancer by Utilizing the Image Pyramids of Whole-Slide Imaging and Multi-scale Convolutional Neural Networks’, in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, Jul. 2019, vol. 1, pp. 696–703. doi: 10.1109/COMPSAC.2019.00105.
[69] J. Tan, Y. Huo, Z. Liang, and L. Li, ‘Expert knowledge-infused deep learning for automatic lung nodule detection’, *Journal of X-Ray Science and Technology*, vol. 27, no. 1, pp. 17–35, Jan. 2019, doi: 10.3233/XST-180426.
[70] H. Feng *et al.*, ‘A knowledge-driven feature learning and integration method for breast cancer diagnosis on multi-sequence MRI’, *Magnetic Resonance Imaging*, vol. 69, pp. 40–48, Jun. 2020, doi: 10.1016/j.mri.2020.03.001.
[71] S. Hochreiter and J. Schmidhuber, ‘Long Short-Term Memory’, *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
[72] A. Yala, C. Lehman, T. Schuster, T. Portnoi, and R. Barzilay, ‘A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction’, *Radiology*, vol. 292, no. 1, pp. 60–66, May 2019, doi: 10.1148/radiol.2019182716.
[73] S. E. Spasov, L. Passamonti, A. Duggento, P. Liò, and N. Toschi, ‘A Multi-modal Convolutional Neural Network Framework for the Prediction of

Alzheimer's Disease', in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2018, pp. 1271–1274. doi: 10.1109/EMBC.2018.8512468.
[74] I. Reda *et al.*, 'Deep Learning Role in Early Diagnosis of Prostate Cancer', *Technol Cancer Res Treat*, vol. 17, p. 1533034618775530, Jan. 2018, doi: 10.1177/1533034618775530.
[75] G. L. Rogova and P. C. Stomper, 'Information fusion approach to microcalcification characterization', *Information Fusion*, vol. 3, no. 2, pp. 91–102, Jun. 2002, doi: 10.1016/S1566-2535(02)00054-4.
[76] D. Lederman, B. Zheng, X. Wang, X. H. Wang, and D. Gur, 'Improving Breast Cancer Risk Stratification Using Resonance-Frequency Electrical Impedance Spectroscopy Through Fusion of Multiple Classifiers', *Ann Biomed Eng*, vol. 39, no. 3, pp. 931–945, Mar. 2011, doi: 10.1007/s10439-010-0210-4.
[77] M. S. B. Sehgal, I. Gondal, and L. Dooley, 'Support vector machine and generalized regression neural network based classification fusion models for cancer diagnosis', in *Fourth International Conference on Hybrid Intelligent Systems (HIS'04)*, Dec. 2004, pp. 49–54. doi: 10.1109/ICHIS.2004.88.
[78] T. Majtner, S. Yildirim-Yayilgan, and J. Y. Hardeberg, 'Combining deep learning and hand-crafted features for skin lesion classification', in *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Dec. 2016, pp. 1–6. doi: 10.1109/IPTA.2016.7821017.
[79] Y. Yoo *et al.*, 'Deep learning of brain lesion patterns and user-defined clinical and MRI features for predicting conversion to multiple sclerosis from clinically isolated syndrome', *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 7, no. 3, pp. 250–259, May 2019, doi: 10.1080/21681163.2017.1356750.
[80] S.-C. Huang, A. Pareek, R. Zamanian, I. Banerjee, and M. P. Lungren, 'Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection', *Scientific Reports*, vol. 10, no. 1, Art. no. 1, Dec. 2020, doi: 10.1038/s41598-020-78888-w.
[81] M. Heath *et al.*, 'Current Status of the Digital Database for Screening Mammography', Springer, Dordrecht, 1998, pp. 457–460. doi: 10.1007/978-94-011-5318-8_75.
[82] M. Heath, K. Bowyer, D. Kopans, R. Moore, and P. K. Jr, 'THE DIGITAL DATABASE FOR SCREENING MAMMOGRAPHY', p. 10.
[83] 'The mini-MIAS database of mammograms'. http://peipa.essex.ac.uk/info/mias.html (accessed Jul. 04, 2021).
[84] W. D. Bidgood Jr, S. C. Horii, F. W. Prior, and D. E. Van Syckle, 'Understanding and Using DICOM, the Data Interchange Standard for Biomedical Imaging', *Journal of the American Medical Informatics Association*, vol. 4, no. 3, pp. 199–212, May 1997, doi: 10.1136/jamia.1997.0040199.
[85] T. F. Chan and L. A. Vese, 'Active contours without edges', *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 266–277, Feb. 2001, doi: 10.1109/83.902291.
[86] H. Høst and E. Lund, 'Age as a prognostic factor in breast cancer', *Cancer*, vol. 57, no. 11, pp. 2217–2221, 1986, doi: 10.1002/1097-0142(19860601)57:11<2217::AID-CNCR2820571124>3.0.CO;2-T.

[87] ‘Breast care essentials: breast calcifications’, *Breast Cancer Now*, Jun. 03, 2015. https://breastcancernow.org/information-support/have-i-got-breast-cancer/benign-breast-conditions/breast-calcifications (accessed May 24, 2021).
[88] ‘Breast imaging reporting and data system (BI-RADS) atlas.’ American College of Radiology, Reston, VA, 2003.
[89] CDCBreastCancer, ‘What Does It Mean to Have Dense Breasts?’, *Centers for Disease Control and Prevention*, Jun. 17, 2020. https://www.cdc.gov/cancer/breast/basic_info/dense-breasts.htm (accessed Jul. 04, 2021).
[90] ‘Breast cancer in women - Causes’, *nhs.uk*, Oct. 24, 2017. https://www.nhs.uk/conditions/breast-cancer/causes/ (accessed May 25, 2021).
[91] A. O. Malagelada, ‘Automatic mass segmentation in mammographic images’, p. 213.
[92] L. W. Bassett, K. Conner, and I. Ms, ‘The Abnormal Mammogram’, *Holland-Frei Cancer Medicine. 6th edition*, 2003, Accessed: Apr. 30, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK12642/
[93] S. Heywang-Koebrunner, I. Schreer, and S. Barter, *Diagnostic Breast Imaging. Mammographym sonography, magnetic resonance imaging and interventional procedures*.
[94] C. Wang, A. R. Brentnall, J. Cuzick, E. F. Harkness, D. G. Evans, and S. Astley, ‘A novel and fully automated mammographic texture analysis for risk prediction: results from two case-control studies’, *Breast Cancer Research*, vol. 19, no. 1, p. 114, Oct. 2017, doi: 10.1186/s13058-017-0906-6.
[95] K. Zuiderveld, ‘Contrast Limited Adaptive Histogram Equalization’, *Graphics Gems*, pp. 474–485, Jan. 1994, doi: 10.1016/B978-0-12-336156-1.50061-6.
[96] P. Lambin *et al.*, ‘Radiomics: Extracting more information from medical images using advanced feature analysis’, *European Journal of Cancer*, vol. 48, no. 4, pp. 441–446, Mar. 2012, doi: 10.1016/j.ejca.2011.11.036.
[97] R. M. Haralick, K. Shanmugam, and I. Dinstein, ‘Textural Features for Image Classification’, *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973, doi: 10.1109/TSMC.1973.4309314.
[98] G. N. Srinivasan and G. Shobha, ‘Statistical Texture Analysis’, in *Proceedings of World Academy of Science, Engineering and Technology*, 2008, pp. 1264–1269.
[99] X. Tang, ‘Texture information in run-length matrices’, *IEEE Transactions on Image Processing*, vol. 7, no. 11, pp. 1602–1609, Nov. 1998, doi: 10.1109/83.725367.
[100] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, ‘Deep Learning to Improve Breast Cancer Detection on Screening Mammography’, *Scientific Reports*, vol. 9, no. 1, Art. no. 1, Aug. 2019, doi: 10.1038/s41598-019-48995-4.
[101] R. Casana Eslava, ‘Identification of Data Structure with Machine Learning: From Fisher to Bayesian networks’, doctoral, Liverpool John Moores University, 2019. doi: 10.24377/LJMU.t.00010869.
[102] W. S. McCulloch and W. Pitts, ‘A logical calculus of the ideas immanent in nervous activity’, *The Bulletin of Mathematical Biophysics*, vol. 5, no. 4, pp. 115–133, Dec. 1943, doi: 10.1007/BF02478259.
[103] R. W. Floyd and R. W., ‘Algorithm 97: Shortest path’, *Communications of the ACM*, vol. 5, no. 6, p. 345, 1962, doi: 10.1145/367766.368168.

[104] S. Warshall and Stephen, ‘A Theorem on Boolean Matrices’, *Journal of the ACM*, vol. 9, no. 1, pp. 11–12, Jan. 1962, doi: 10.1145/321105.321107.
[105] Y. Takane, F. W. Young, and J. de Leeuw, ‘Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features’, *Psychometrika*, vol. 42, no. 1, pp. 7–67, Mar. 1977, doi: 10.1007/BF02293745.
[106] M. Srivastava, I. Olier, P. Riley, P. Lisboa, and S. Ortega-Martorell, ‘Classifying and Grouping Mammography Images into Communities Using Fisher Information Networks to Assist the Diagnosis of Breast Cancer’, Springer, Cham, 2020, pp. 304–313. doi: 10.1007/978-3-030-19642-4_30.
[107] W. Xu, W. Liu, L. Li, G. Shao, and J. Zhang, ‘Identification of Masses and Microcalcifications in the Mammograms Based on Three Neural Networks: Comparison and Discussion’, in *2008 2nd International Conference on Bioinformatics and Biomedical Engineering*, May 2008, pp. 2299–2302. doi: 10.1109/ICBBE.2008.907.
[108] C. Abirami, R. Harikumar, and S. R. S. Chakravarthy, ‘Performance analysis and detection of micro calcification in digital mammograms using wavelet features’, in *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, Mar. 2016, pp. 2327–2331. doi: 10.1109/WiSPNET.2016.7566558.
[109] M. A. Mazurowski, J. Y. Lo, B. P. Harrawood, and G. D. Tourassi, ‘Mutual information-based template matching scheme for detection of breast masses: From mammography to digital breast tomosynthesis’, *Journal of Biomedical Informatics*, vol. 44, no. 5, pp. 815–823, Oct. 2011, doi: 10.1016/J.JBI.2011.04.008.
[110] M. Elter and E. Haßlmeyer, ‘A knowledge-based approach to the CADx of mammographic masses’, Mar. 2008, vol. 6915, p. 69150L. doi: 10.1117/12.770135.
[111] C.-H. Wei, Y. Li, and P. J. Huang, ‘Mammogram retrieval through machine learning within BI-RADS standards’, *Journal of Biomedical Informatics*, vol. 44, no. 4, pp. 607–614, Aug. 2011, doi: 10.1016/J.JBI.2011.01.012.
[112] R. G. Raidou *et al.*, ‘Visual Analytics for the Exploration of Tumor Tissue Characterization’, *Computer Graphics Forum*, vol. 34, no. 3, pp. 11–20, 2015, doi: 10.1111/cgf.12613.
[113] H. Ruiz, S. Ortega-Martorell, I. H. Jarman, J. D. Martín, and P. J. G. Lisboa, ‘Constructing similarity networks using the Fisher information metric’, in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2012, pp. 191–6.
[114] H. Ruiz, I. H. Jarman, J. D. Martín, and P. J. G. Lisboa, ‘The role of Fisher information in primary data space for neighbourhood mapping.’, in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2011, pp. 381–6.
[115] S. M. Pizer *et al.*, ‘Adaptive histogram equalization and its variations’, *Computer Vision, Graphics, and Image Processing*, vol. 39, no. 3, pp. 355–368, Sep. 1987, doi: 10.1016/S0734-189X(87)80186-X.
[116] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, ‘Learning representations by back-propagating errors’, *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, doi: 10.1038/323533a0.
[117] S. Haykin, *Neural networks: a comprehensive foundation*, 2nd ed. Prentice Hall, 1998.

[118] I. Borg and P. J. F. Groenen, *Modern multidimensional scaling*. New York: Springer Verlag, 1997.
[119] K. Zuiderveld, 'Contrast Limited Adaptive Histogram Equalization', *Graphics Gems*, pp. 474–485, Jan. 1994, doi: 10.1016/B978-0-12-336156-1.50061-6.
[120] R. M. Haralick, K. Shanmugam, and I. Dinstein, 'Textural Features for Image Classification', *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973, doi: 10.1109/TSMC.1973.4309314.
[121] G. N. Srinivasan and G. Shobha, 'Statistical Texture Analysis', in *Proceedings of World Academy of Science, Engineering and Technology*, 2008, pp. 1264–1269.
[122] L. Breiman, 'Random Forests', *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[123] P. J. G. Lisboa, T. A. Etchells, I. H. Jarman, and S. J. Chambers, 'Finding reproducible cluster partitions for the k-means algorithm', *BMC bioinformatics*, vol. 14 Suppl 1, no. Suppl 1, p. S8, Jan. 2013, doi: 10.1186/1471-2105-14-S1-S8.
[124] H.-H. Bock, 'Origins and extensions of the k-means algorithm in cluster analysis', *Electronic journal for history of probability and statistics*, vol. 4, no. 2, pp. 1–18, 2008.
[125] C. Parmar *et al.*, 'Robust Radiomics Feature Quantification Using Semiautomatic Volumetric Segmentation', *PLoS ONE*, vol. 9, no. 7, p. e102107, Jul. 2014, doi: 10.1371/journal.pone.0102107.
[126] L. Shen, *lishen/end2end-all-conv*. 2021. Accessed: May 02, 2021. [Online]. Available: https://github.com/lishen/end2end-all-conv
[127] P. Xi, C. Shu, and R. Goubran, 'Abnormality Detection in Mammography using Deep Convolutional Neural Networks', in *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, Jun. 2018, pp. 1–6. doi: 10.1109/MeMeA.2018.8438639.
[128] P. Xi, H. Guan, C. Shu, L. Borgeat, and R. Goubran, 'An integrated approach for medical abnormality detection using deep patch convolutional neural networks', *Vis Comput*, vol. 36, no. 9, pp. 1869–1882, Sep. 2020, doi: 10.1007/s00371-019-01775-7.
[129] E. Rashed and M. S. A. El Seoud, 'Deep learning approach for breast cancer diagnosis', in *Proceedings of the 2019 8th International Conference on Software and Information Engineering - ICSIE '19*, New York, New York, USA, 2019, pp. 243–247. doi: 10.1145/3328833.3328867.
[130] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition', in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.
[131] S. Lee, M. Amgad, M. Masoud, R. Subramanian, D. Gutman, and L. Cooper, 'An Ensemble-based Active Learning for Breast Cancer Classification', in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Nov. 2019, pp. 2549–2553. doi: 10.1109/BIBM47256.2019.8983317.
[132] P. J. G. Lisboa, S. Ortega-Martorell, and I. Olier, 'Explaining the Neural Network: A Case Study to Model the Incidence of Cervical Cancer', in *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Cham, 2020, pp. 585–598. doi: 10.1007/978-3-030-50146-4_43.
[133] I. Olier, A. Sansom, P. Lisboa, and S. Ortega-Martorell, 'Using MLP partial responses to explain in-hospital mortality in ICU', in *2020 International Conference on Data Analytics for Business and Industry: Way Towards a*

Sustainable Economy (ICDABI), Oct. 2020, pp. 1–5. doi: 10.1109/ICDABI51230.2020.9325691.
[134] P. J. G. Lisboa, S. Ortega-Martorell, S. Cashman, and I. Olier, 'The Partial Response Network: a neural network nomogram', *arXiv:1908.05978 [cs, stat]*, Jun. 2020, Accessed: May 13, 2021. [Online]. Available: http://arxiv.org/abs/1908.05978
[135] G. Hooker, 'Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables', *Journal of Computational and Graphical Statistics*, vol. 16, no. 3, pp. 709–732, Sep. 2007, doi: 10.1198/106186007X237892.
[136] J. Friedman, T. Hastie, and R. Tibshirani, 'Regularization Paths for Generalized Linear Models via Coordinate Descent', *J Stat Softw*, vol. 33, no. 1, pp. 1–22, 2010.
[137] D. Krstajic, L. J. Buturovic, D. E. Leahy, and S. Thomas, 'Cross-validation pitfalls when selecting and assessing regression and classification models', *J Cheminform*, vol. 6, p. 10, Mar. 2014, doi: 10.1186/1758-2946-6-10.
[138] R. M. Rangayyan, N. R. Mudigonda, and J. E. L. Desautels, 'Boundary modelling and shape analysis methods for classification of mammographic masses', *Med. Biol. Eng. Comput.*, vol. 38, no. 5, pp. 487–496, Sep. 2000, doi: 10.1007/BF02345742.
[139] M. S. Newell, *Round and Punctate Calcifications*. Oxford University Press. Accessed: May 25, 2021. [Online]. Available: https://oxfordmedicine.com/view/10.1093/med/9780190270261.001.0001/med-9780190270261-chapter-37
[140] P. L. Arancibia Hernández, T. Taub Estrada, A. López Pizarro, M. L. Díaz Cisternas, and C. Sáez Tapia, 'Calcificaciones mamarias: descripción y clasificación según la 5.a edición BI-RADS', *Revista Chilena de Radiología*, vol. 22, no. 2, pp. 80–91, Apr. 2016, doi: 10.1016/j.rchira.2016.06.004.
[141] N. Bhagwat, J. D. Viviano, A. N. Voineskos, M. M. Chakravarty, and A. D. N. Initiative, 'Modeling and prediction of clinical symptom trajectories in Alzheimer's disease using longitudinal data', *PLOS Computational Biology*, vol. 14, no. 9, p. e1006376, Sep. 2018, doi: 10.1371/journal.pcbi.1006376.
[142] J. Yap, W. Yolland, and P. Tschandl, 'Multimodal skin lesion classification using deep learning', *Experimental Dermatology*, vol. 27, no. 11, pp. 1261–1267, 2018, doi: https://doi.org/10.1111/exd.13777.
[143] R. Yan *et al.*, 'Richer fusion network for breast cancer classification based on multimodal data', *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, p. 134, Apr. 2021, doi: 10.1186/s12911-020-01340-6.
[144] D. Ramachandram and G. W. Taylor, 'Deep Multimodal Learning: A Survey on Recent Advances and Trends', *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, Nov. 2017, doi: 10.1109/MSP.2017.2738401.
[145] T. L. R. Medicine, 'Opening the black box of machine learning', *The Lancet Respiratory Medicine*, vol. 6, no. 11, p. 801, Nov. 2018, doi: 10.1016/S2213-2600(18)30425-9.
[146] Cancer Research UK, 'Prostate cancer statistics | Cancer Research UK'. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/prostate-cancer#heading-Zero (accessed Nov. 19, 2019).
[147] National Collaborating Centre for Cancer (UK), 'Prostate cancer: diagnosis and treatment', Cardiff, 2014.

[148] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman, ‘Computer-Aided Detection of Prostate Cancer in MRI’, *IEEE Transactions on Medical Imaging*, vol. 33, no. 5, pp. 1083–1092, 2014, doi: 10.1109/TMI.2014.2303821.
[149] S. Liu, H. Zheng, Y. Feng, and W. Li, ‘Prostate Cancer Diagnosis using Deep Learning with 3D Multiparametric MRI’, Mar. 2017. doi: 10.1117/12.2277121.
[150] A. Mehrtash *et al.*, ‘Classification of Clinical Significance of MRI Prostate Findings Using 3D Convolutional Neural Networks’, in *SPIE Medical Imaging 2017: Computer-Aided Diagnosis*, 2017. doi: 10.1117/12.2277123.
[151] A. Kitchen and J. Seah, ‘Support vector machines for prostate lesion classification’, in *SPIE Medical Imaging 2017: Computer-Aided Diagnosis*, Mar. 2017, vol. 10134, p. 1013427. doi: 10.1117/12.2277120.
[152] J. Gallagher, ‘Prostate cancer treatment “not always needed” - BBC News’. https://www.bbc.co.uk/news/health-37362572 (accessed Dec. 01, 2018).
[153] F. Brimo *et al.*, ‘Contemporary Grading for Prostate Cancer: Implications for Patient Care’, *European Urology*, vol. 63, no. 5, pp. 892–901, May 2013, doi: 10.1016/J.EURURO.2012.10.015.
[154] D. L. Langer *et al.*, ‘Prostate Tissue Composition and MR Measurements: Investigating the Relationships between ADC, T2, Ktrans, ve, and Corresponding Histologic Features’, *Radiology*, vol. 255, no. 2, pp. 485–494, 2010, doi: 10.1148/radiol.10091343.
[155] C. M. Bishop, M. Svensén, and C. K. I. Williams, ‘The Generative Topographic Mapping’, 1998.
[156] T. Kohonen, *Self-Organizing Maps*, vol. 30. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001. doi: 10.1007/978-3-642-56927-2.
[157] ‘AI “outperforms” doctors diagnosing breast cancer’, *BBC News*, Jan. 02, 2020. Accessed: Aug. 14, 2021. [Online]. Available: https://www.bbc.com/news/health-50857759
[158] ‘Breast Screening Programme, England 2019-20’, *NHS Digital*. https://digital.nhs.uk/data-and-information/publications/statistical/breast-screening-programme/england---2019-20 (accessed Aug. 14, 2021).