School of Computing, Engineering and Digital Technologies
Department of Computing and Games
Teesside University
Middlesbrough TS1 3BA

Improving Classification of Response to Neoadjuvant Therapy in Breast Cancer Patients Through the Late Fusion of Clinical and DCE-MRI Data

An academic research paper
Submitted in partial requirements for the degree of MSc Applied Artificial Intelligence

Date: 18/08/2022

B1414659
Catherine Vaughan-Jackson

Supervisor: Olugbnga Akinade

Acknowledgements

I would like to thank my supervisor Olugbnga Akinade for his guidance through this module and the insight he has been able to provide. I would also like to express my utmost gratitude to Annalisa Occhipinti for her constant support and knowledge. I would also like to thank Alessandro Di Stefano and Teesside University for the excellent support and resources provided for this whole course.

Contents

Introduction

Methodology

Results and Discussion

Conclusions

References

Improving Classification of Response to Neoadjuvant Therapy in Breast Cancer Patients Through the Late Fusion of Clinical and DCE-MRI Data

Abstract

A paper published by the British medical journal by Vaidya et al (2018) suggests that neoadjuvant chemotherapy may not be beneficial to all patients, and we should consider reducing the widespread use of neoadjuvant chemotherapy. Utilising machine learning, we can create a tool to aid the decision process of physicians by classifying individuals who would not benefit from this form of treatment. Due to the complex nature in predicting patients who have response to neoadjuvant therapy accuracy has been limited. Therefore, we propose the use of late fusion of multi-modal data using weighted averages to improve classification accuracy. In this paper we outline the steps to create an end-to-end pipeline for fusing image and clinical data at the decision level. Despite limitations in our dataset, we have successfully seen an increase from 76% accuracy and 77% F1 score in our clinical model and 59% accuracy 70% F1 in our image model to 82% accuracy and a F1 score of 84%. In addition to this, when applying the same fusion pipeline with recurrence of cancer as the target variable we have achieved an increase from 71% accuracy in the clinical data, 67% in the image data to 80% through late fusion.

Keywords: Deep Learning, Convolutional Neural Networks, Multi-modal, Late Fusion, Neoadjuvant Therapy, Classification

Introduction and Background

Breast cancer is the second most common malignancy in women worldwide, about one in eight women will be diagnosed with it at some point in her life (Society, 2020). According to Breast Cancer UK, Breast cancer is the fourth most common cause of cancer death and the second most common cause of any death in women (BCUK, 2022). Despite this, the UK employs an extensive screening programme which has contributed to an improved Five-year survival rate reaching 85.6% in 2010-2014 (Nuffield, 2021) and since the mid-1980s, breast cancer mortality rates have decreased by 45% (BCUK, 2022). The early stages in the diagnosis and treatment of cancer are vital for long term survival with advancements within artificial intelligence and machine learning offering the possibility of developing highly accurate models. There has already been extensive research within the area of diagnosis. For example, Prakash and Visakha (2020) use neural networks to predict malignancy in breast tissue with an F1 score of 98% (Wisconsin UCL dataset) and Xie et al. (2019) use histopathological slide images to classify breast cancer with an AUC score of 97%. These discoveries have been put into practice Leibig et al (2022) categorised cancers into three categories from mammograms as “confident normal,” “not confident” (in which no prediction is given), and “confident cancer.” The doctor and AI working together was 2.6% better at detecting breast cancer than a doctor working alone and raised fewer false alarms. It accomplished this while automatically setting aside scans it classified as “confident normal” which amounted to 63% of all mammograms reducing workload considerably.

In comparison, there has been less research pertaining to the treatment process of a patient. The process through which an individual is given a formal diagnosis and prognosis is complex and can involve multiple scans and biopsies and both a physician and pathologist must come to agreement. Treatment plans are unique to each individual and professional opinions can often differ on course of action, particularly with those for those with the most invasive forms of cancer. By utilising the power of machine learning we can process vast quantities of data over many different parameters to find trends which might help aid the physician's decision. An area of particular interest is aiding the decision of whether a patient should undergo neoadjuvant treatment. An article “Rethinking neoadjuvant chemotherapy for breast cancer” by Vaidya et al. (2018) raises concerns about the use of neoadjuvant treatment, acknowledging that neoadjuvant chemotherapy may not be beneficial to patients and even suggests reducing the widespread use of neoadjuvant chemotherapy. As stated, “Neoadjuvant chemotherapy is being increasingly used for breast cancer despite higher rates of local recurrence and no evidence of survival benefit, mainly because of the immediate and dramatic pathological responses seen with newer drugs”. Therefore, if we can use machine learning to predict those who would not benefit from this form of treatment, we could advance to surgery more quickly for those individuals and potentially improve the survival rate.

The aim of neoadjuvant chemotherapy is to shrink a tumour or stop the spread of cancer to make surgery less invasive and more effective. Usually there are several factors that are considered to decide if a person should undergo neoadjuvant treatment, including:

- Cancer type and stage
- Whether or not the cancer has spread to lymph nodes
- The goal of treatment, whether it is to rid your body of cancer, slow the cancer’s growth and progression, or ease the symptoms of your cancer
- How well your body is likely to tolerate multiple treatments

Although these are usually the main deciding factors, there are potentially many more variables that can contribute to determine an individual's response. For example, ER/PR/HER2 positivity, genetic history, own personal medical history, genetics and even potentially information found in imagery. As previously stated, developments in deep learning grant us the ability to analyse and process vast quantities of data and process complex image data enabling us to build accurate predictive models that can potentially find connections between features that are not currently considered in the decision process. The complex nature of the factors involved in response to neoadjuvant treatment has consequently seen a limitation in accuracy in predictive models compared to diagnostic models. For example, Ha et. al (2019) research the use of DCNN Prior to Initiation of Chemotherapy to predict breast tumour response achieving around 87% for the three response categories. Chen et al. 2020 uses machine learning to predict patients who have a pathological complete response to neoadjuvant chemotherapy achieving AUC 0.834, specificity 73.21%, and sensitivity 80%.

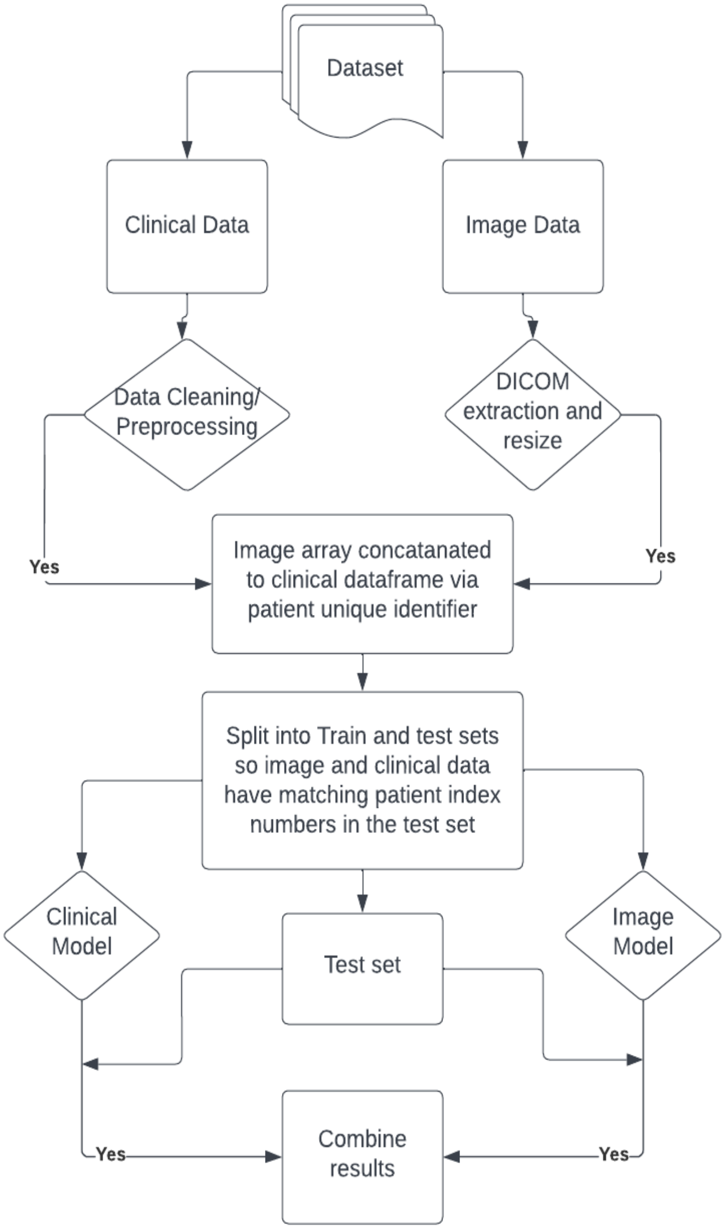
Proposed Solution

For this paper we propose the utilisation of late multimodal data integration, using both DCE-MR image and clinical data to enable a more rounded perspective on a patient and therefore a more accurate model. In this paper we will address the following:

- The efficacy of late multimodal integration for predicting response to neoadjuvant treatment
- What features contribute to a positive prediction

There are three methods for data fusion: early, joint, and late. Early fusion (feature level fusion) integrates multiple input modalities into a single feature vector before being input into the deep learning model. Inputs can be fused using original features or extracted features. The joint (or mid-level) fusion integrates during the intermediate layers of a neural network The key difference, compared to early fusion, is that the loss is propagated back to the feature extracting neural networks during training, thus creating better feature representations for each training iteration. The late (or decision-level) fusion trains modalities separately and aggregates the individual decisions to obtain the final prediction. This can be achieved using several different methods such as max-fusion, average-fusion, or Bayes rules. Although the late fusion ignores some low-level interactions between modalities, it allows easy training with more flexibility, and simplicity which is particularly useful when dealing with large datasets. For this paper we will focus on the efficacy of late fusion techniques, figure 1 outlines the processes undertaken to achieve this.

Figure 1. Depicts the pipeline for our late fusion model, the clinical data is cleaned, and image data is converted into a numpy array. The image array is converted into a dataframe and merged on the patient ID number with the clinical data. The merged data frame is then split into train and test sets and split back into image and clinical sets. Once both models are run, a predictive value for each patient in the test set for each model is evaluated. A weighted average combines the predictive values to produce our new prediction.


Dataset
Clinical Data
Image Data
Data Cleaning/ Preprocessing
DICOM extraction and resize
Yes
Yes
Image array concatanated to clinical dataframe via patient unique identifier
Split into Train and test sets so image and clinical data have matching patient index numbers in the test set
Clinical Model
Test set
Image Model
Yes
Combine results
Yes

Literature Review

Multi-modal data integration

Xie et al. (2019) proposed a deep learning method to classify 14 bird sounds by extracting acoustic features and time frequency visual features then using a classifier model on the acoustic features, a CNN on the time frequency visuals and then further extracting features from the visuals and classifying. All three models are then combined to create a predictive result which achieved the best F1-score 94.36%, which is higher than using the acoustic features approach (88.97%) and using the visual features approach (88.87%). Another Paper by Pandera et al. (2021) explores late fusion for emotion classification of music videos and achieves 88.56% in accuracy, 0.88% in F1-score, and 0.987% AUC score.

Specifically referring to late fusion in medical data, a review paper by Huang et al. (2020) explores in detail early, joint and late fusion methods with medical imaging and clinical data across 17 different research papers with various levels of success. They find that multimodality fusion models generally led to increased accuracy (1.2–27.7%) and AUROC (0.02–0.16) over traditional single modality models for the same task. But determined that, no single fusion strategy consistently led to optimal performance across all domains.

Yala Et.al (2019) use a joint fusion model to predict the breast cancer risk within 5 years using risk factor features for the EHR and mammography imagery to achieve a significantly higher AUC (0.70) compared to DL image classification (0.62; P , .001) and Risk Factor Logistic regression model (0.67; P = .01). Silva et al (2021) develop MultiSurv, a multimodal deep learning method for long-term pan-cancer survival prediction MultiSurv was applied to data from 33 different cancer types and yields accurate pan-cancer patient survival curves. The imaging model was trained on weights from a ResNeXt-50 convolutional neural network(pre-trained on imagenet and natural image dataset) and the five other models were trained with fully connected neural networks with up to five hidden layers. Different combinations of modalities were tested and found clinical data with mRNA data achieved a C-index of 0.822 (0.805–0.837).

Hao et.al (2020) propose PAGE-Net, a model that integrates histopathological images and genomic data, not only to improve survival prediction, but also to identify genetic and histopathological patterns that cause different survival rates in patients. PAGE-Net achieved a C-index of 0.702, which is higher than the results achieved with only histopathological images (0.509).

Methodology

Dataset Description

Despite great advancements within the machine learning community, access to suitable datasets for experimentation and research are limited, particularly within the medical field. Due to ethical clearance, privacy and time constraints to collate a dataset from scratch for this research is not possible. The most extensive, thorough dataset found containing medical image and clinical data is the Duke breast cancer dataset (Saha et. Al., 2018), and is freely available open source. Initial researchers (Saha et al., 2018) used radiomics (the extraction of image features) to explore correlations between imaging characteristics and its genomic composition. Each distinct molecular subtype is associated with a particular tendency of disease progression, therefore identification of those subtypes is vital for a physician's treatment plan. 529 features were extracted and input into machine learning classification models to predict biometric markers of an individual patient. Results of these experiments were predictive of Luminal A subtype with AUC = 0.697 (95% CI: 0.647–0.746, $p < .0001$), triple negative breast cancer with AUC = 0.654 (95% CI: 0.589–0.727, $p < .0001$), ER status with AUC = 0.649 (95% CI: 0.591–0.705, $p < .001$), and PR status with AUC = 0.622. Other than this initial research there have not been any published works using this data.

This dataset consists of 922 patients, each patient has corresponding anonymised clinical and MRI image data. For each patient there are four to six folders containing volumetric MRI scans in DICOM format: a non-fat saturated T1-weighted sequence, a fat-saturated gradient echo T1-weighted pre-contrast sequence, and mostly three to four post-contrast sequences (DCE-MRI). The clinical data consists of 91 different parameters which are listed in Figure 2. The demographic for this data consists entirely of women with malignant cancer and ages range between 21 and 89 years with a mean of 53 years.

For our target variable of Response to Neoadjuvant treatment there are unfortunately only 256 patients who undertook neoadjuvant treatment considerably reducing our dataset size. Therefore, we must take into consideration that our results will be affected by the limitation of data and our models are at high risk of overfitting and poor accuracy. Therefore, in our experimentation stage we will compare how our proposed concept compares when using “Recurrence Event(s)” as our target variable, although we can use nearly the whole dataset, this variable is highly imbalanced (87 positive and 822 negative) and therefore comes with its own limitations.

Figure 2.

Duke Breast Cancer Dataset (922 patients)

Column types	**Number of columns**	**Column names**	**Comments**
Patient ID	1	‘Patient ID’	
MRI Technical Information	18	‘Days to MRI (From the Date of Diagnosis)’, ‘Manufacturer’, ‘Manufacturer Model Name’, ‘Scan Options’,'Field Strength (Tesla)', 'Patient Position During MRI', 'Image Position of Patient', 'Contrast Agent','Contrast Bolus Volume (mL)', 'TR (Repetition Time)', 'TE (Echo Time)', 'Acquisition Matrix', 'Slice Thickness ', 'Rows', 'Columns', 'Reconstruction Diameter ', 'Flip Angle \n', 'FOV Computed (Field of View) in cm '	
Demographics	4	'Date of Birth (Days)', 'Menopause (at diagnosis)', 'Race and Ethnicity', 'Metastatic at Presentation (Outside of Lymph Nodes)’	
Tumour Characteristics	15	'ER', 'PR', 'HER2', 'Mol Subtype', 'Oncotype score', 'Staging(Tumor Size)# [T]', 'Staging(Nodes)#(Nx replaced by -1)[N]', 'Staging(Metastasis)#(Mx -replaced by -1)[M]', 'Tumor Grade(T)', Tumor Grade(N)', Tumor Grade(M)', 'Nottingham grade', 'Histologic type', 'Tumor Location', 'Position', 'Bilateral Information', 'If Bilateral, Different Rec Status', ‘Side Annotated', 'For Other Side If Bilateral',’ Oncotype score (for the other side){##}', ‘Nottingham grade (for the other side){#}', 'ER (for the other side)', 'PR (for the other side)', 'HER2 (for the other side)', ‘Mol Subtype (for the other side)’	
MRI Findings	5	'Multicentric/Multifocal','Contralateral Breast Involvement','Lymphadenopathy or Suspicious Nodes', 'Skin/Nipple Involvement', 'Pec/Chest Involvement	Described as not reliable
Surgery	3	Surgery','Days to Surgery (from the date of diagnosis)', 'Definitive Surgery Type	
Radiation Therapy	2	Neoadjuvant Radiation Therapy', 'Adjuvant Radiation Therapy	
Chemotherapy	2	Neoadjuvant Chemotherapy Therapy', 'Adjuvant Chemotherapy Therapy	
Hormone/Al/Herceptin therapy	2	Neoadjuvant Hormone/Al/Herceptin Therapy', 'Adjuvant Hormone/Al/Herceptin Therapy	
Tumour response	2	Clinical Response, Evaluated Through Imaging ', 'Pathologic Response to Neoadjuvant Therapy'	

Recurrence	3	'Recurrence event(s)','Days to local recurrence (from the date of diagnosis) ', 'Days to distant recurrence(from the date of diagnosis) ', 'Days to death (from the date of diagnosis) ', 'Days to last local recurrence free assessment (from the date of diagnosis) ', 'Days to last distant recurrence free assessment(from the date of diagnosis)	
Follow Up	4	Age at last contact in EMR f/u(days)(from the date of diagnosis) ,last time patient known to be alive, unless age of death is reported(in such case the age of death)	
Mammogram	8	'Age at mammo (days)', 'Breast Density', 'Shape', 'Margin', 'Architectural distortion', 'Mass Density', 'Calcifications', 'Tumor Size (cm)'	Majority NA
US features	6	'Shape.1', 'Margin.1', 'Tumor Size (cm).1', 'Echogenicity', 'Solid', 'Posterior acoustic shadowing'	Majority NA
Chemotherapy	2	Neoadjuvant Chemotherapy', 'Adjuvant Chemotherapy'	
Endocrine Therapy	5	Neoadjuvant Endocrine Therapy Medications ', 'Adjuvant Endocrine Therapy Medications ', 'Known Ovarian Status ', 'Number of Ovaries In Situ \n', 'Therapeutic or Prophylactic Oophorectomy as part of Endocrine Therapy ',Neoadjuvant Endocrine Therapy Medications ', 'Adjuvant Endocrine Therapy Medications ', 'Known Ovarian Status ', 'Number of Ovaries In Situ \n', 'Therapeutic or Prophylactic Oophorectomy as part of Endocrine Therapy ',	
Anti-Her2 Neu Therapy	2	'Neoadjuvant Anti-Her2 Neu Therapy', 'Adjuvant Anti-Her2 Neu Therapy ',	
Neoadjuvant therapy	4	'Received Neoadjuvant Therapy or Not', 'Pathologic response to Neoadjuvant therapy: Pathologic stage (T) following neoadjuvant therapy ', 'Pathologic response to Neoadjuvant therapy: Pathologic stage (N) following neoadjuvant therapy', 'Pathologic response to Neoadjuvant therapy: Pathologic stage (M) following neoadjuvant therapy '	
Near complete response	3	'Overall Near-complete Response: Stricter Definition', 'Overall Near-complete Response: Looser Definition', 'Near-complete Response (Graded Measure)'	

Data Acquisition

All 922 patient files (400GB) were downloaded using NBIA data retriever. There are variations in how folders are named and how many different types of MRI contrast images each patient had. Most had a fat saturated scan, non fat saturated scan and three to six different contrasts. After carefully viewing batches of MRI scans at different contrasts using RadiAnt viewer, the second non-fat contrast images are the clearest and most consistent across the various types of MRI machines. Code was then written to extract each uniquely named folder of DICOM files and rename as the correct patient identifier narrowing the dataset to 80GB.

Clinical Data Pre-processing and Cleaning

To understand the steps taken to clean the data, we will explain what each feature means and its relation to our target variable. We will also consider the distribution of values for each feature as this will contribute to removal of categories or NA values if highly imbalanced. The category “Response to treatment” is a feature extracted column created by the combination of 'Staging(Tumor Size)# [T]' (T_1), 'Staging(Nodes)#(Nx replaced by -1)[N]' (N_1), 'Pathologic response to Neoadjuvant therapy: Pathologic stage (T) following neoadjuvant therapy ' (T_2) and 'Pathologic response to Neoadjuvant therapy: Pathologic stage’ (N_2) following neoadjuvant therapy'. For this paper we are combining partial and complete response as any positive reduction in tumour is a beneficial reason to have neoadjuvant treatment. Response to treatment is positive where following equations are satisfied:

$$T_2 - T_1 < 0 \text{ and } T_1 > 0$$
$$\text{or}$$
$$N_2 - N_1 < 0 \text{ and } N_1 > 0$$

Tumour Staging	**Number of entries**		**Number of entries with change in tumour grading after neoadjuvant treatment**		
	Before Neoadjuvant treatment	After Neoadjuvant treatment	Complete or partial response	No response	Advancement of cancer
Tumour Size (T)	916	228	81	96	10
Lymph Nodes (N)	898	204	35	138	14
Metastasis (M)	699	5	–removed, not enough instances		

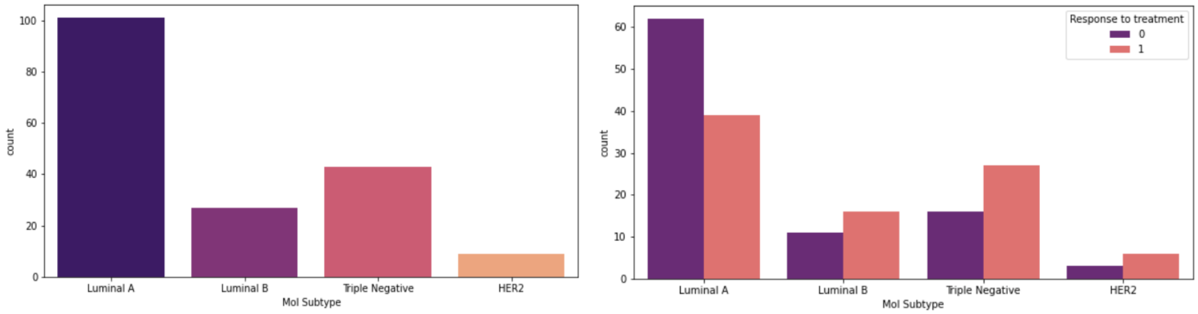
Molecular subtype

There are four breast cancer subtypes defined by immunohistochemistry and these are defined by the expression of estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2)

- Group 1 (luminal A). This group includes tumours that are ER positive and PR positive, but negative for HER2.
- Group 2 (luminal B). This type includes tumours that are ER positive, PR negative and HER2 positive.
- Group 3 (HER2 positive). This type includes tumours that are ER negative and PR negative, but HER2 positive.
- Group 4 (basal-like). This type, which is also called triple-negative breast cancer, includes tumours that are ER negative, PR negative and HER2 negative.

A paper by Onlito et al. (2009) compared outcomes for each of the four subtypes and found that the triple negative subtype has the worst overall survival and disease-free survival therefore these are the most at-risk patients. Below is a chart showing how each of the four subtypes are distributed in our data. We can see in figure 3 that a high proportion of the data is of Luminal A type but in terms of response to treatment, triple negative type patients were the most responsive to treatment with almost double the number of patients having a response compared to not. Similarly, HER2 positive patients were also quite responsive to treatment although a very small number of instances in our dataset.

Figure 3. Distribution of Molecular Subtype



.

Staging of Tumour

Staging Type	**Label**	**Meaning**
Tumour Size (T)	TX	Main tumour cannot be measured
	T0	Main tumour cannot be found
	T1,T2,T3,T4	Refers to the size and/or extent of the main tumour, the higher the number after the T, the larger the tumour or the more it has grown into nearby tissues.
Regional Lymph Nodes (N)	NX	Cancer in nearby lymph nodes cannot be measured
	N0	There is no cancer in nearby lymph nodes
	N1,N2,N3	Refers to the number and location of lymph nodes that contain cancer. The higher the number after the N, the more lymph nodes that contain cancer.
Distant Metastasis	MX	Metastasis cannot be measured
	M0	Cancer has not spread to other parts of the body
	M1	Cancer has spread to other parts of the body

The values described as NX or MX are NA values as there is no way to evaluate the current state of the tumour(s). Also the category “Metastasis” is removed from the analysis, as although very insightful to the severity of cancer after removing NA values we lost a considerable amount of patients from the sample size, and in addition there were only a few cases of those with metastasis therefore would not be a powerful determinant for our deep learning model to learn from.

Below are Figure 4, 5 and 6. Figure 4 shows the distribution of each staging in our data, Showing tumour grade 2 to be the most common and grade 1 staging nodes. In Figure 5 this is elaborated upon by splitting the staging types into those who have response to treatment. To strongly emphasise the impact of each grading, figure 6 represents how each grading is distributed in terms of response by rebalancing each grade proportionally (e.g. a full bar represent 100% of that grading), and we can see about 90% of the cases who have grade 1 tumour size do not respond to neoadjuvant treatment. This is particularly insightful suggesting the larger the tumour the more likely a patient will respond to treatment; on the other hand node grading doesn’t seem to infer a trend as a univariate.

Figure 4. Count of patients for each grading for tumour size and nodes.

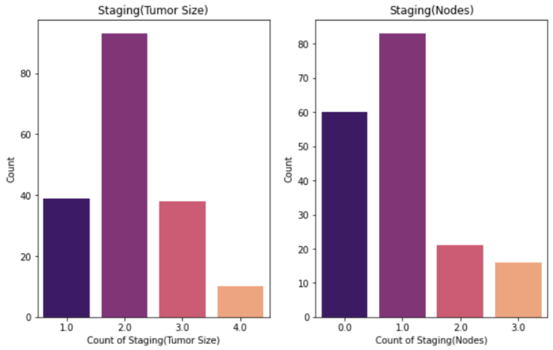


Figure 5. Count of patients for each grading for tumour size and nodes categorised by response to neoadjuvant treatment

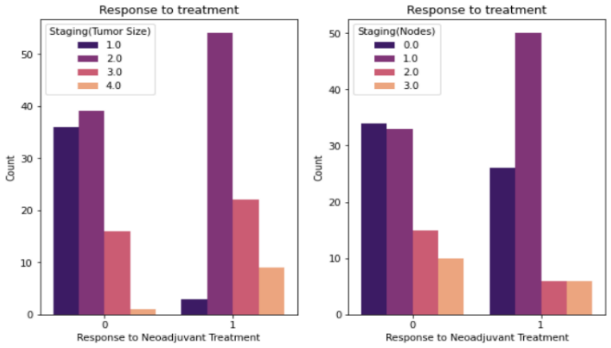
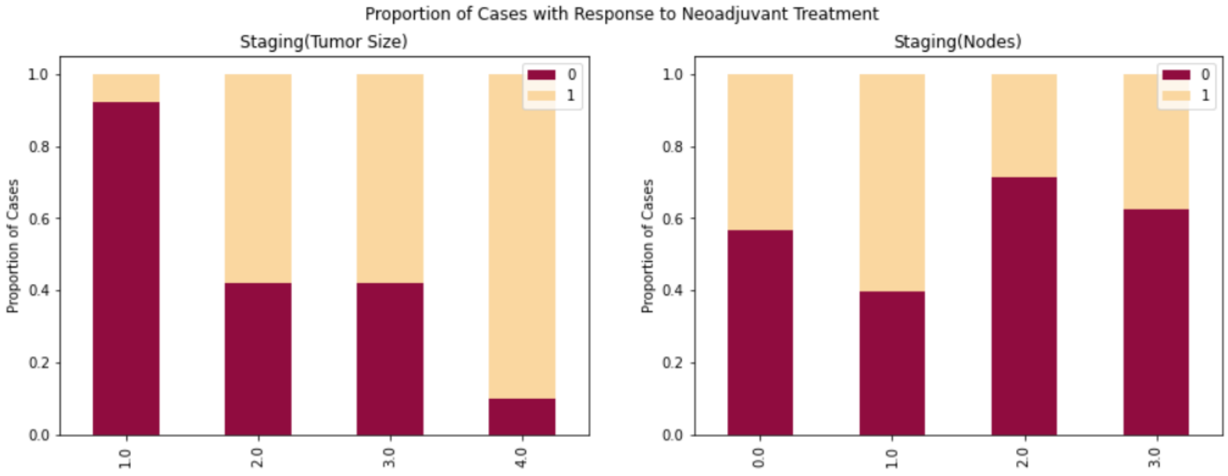


Figure 6. Proportional representation of each grading of tumour stage, split by response to treatment. (1 = full or partial response, 0 = no response)



Tumour type

In biopsy we determine the aggressiveness of the cancer cells; these are defined as three categories: nuclear grade, mitotic rate and tubule formations.

- Nuclear Grade: A score is given from 1 to 3, based on what the nucleus of the cancer cells looks like compared to normal cells. In nuclear grade 1, the nucleus of the cancer cells looks more like normal cells, while in nuclear grade 3, it looks the least like normal cells.

- Mitotic Rate: Describes how quickly the cancer cells are multiplying or dividing, 1 being the slowest, 3 the quickest.
- Tubule formation: This score represents the percent of cancer cells that are formed into tubules. A score of 1 means more than 75% of cells are in tubule formation. A score of 2 is between 10 and 75%. A score of 3 is used when less than 10% of cells are in tubule formation.

Figure 7 and figure 8 below show the count of each tumour type and how each grade is represented proportionally. From the redistributed data we can see that as the nuclear and mitotic grade increase so does the number of cases of patients that have had response to treatment. Although tubular grading doesn't show any obvious trends this might be because most cases are grade 3.

Figure 6. Distribution of grading for Tubule, Nuclear and Mitotic Tumour types

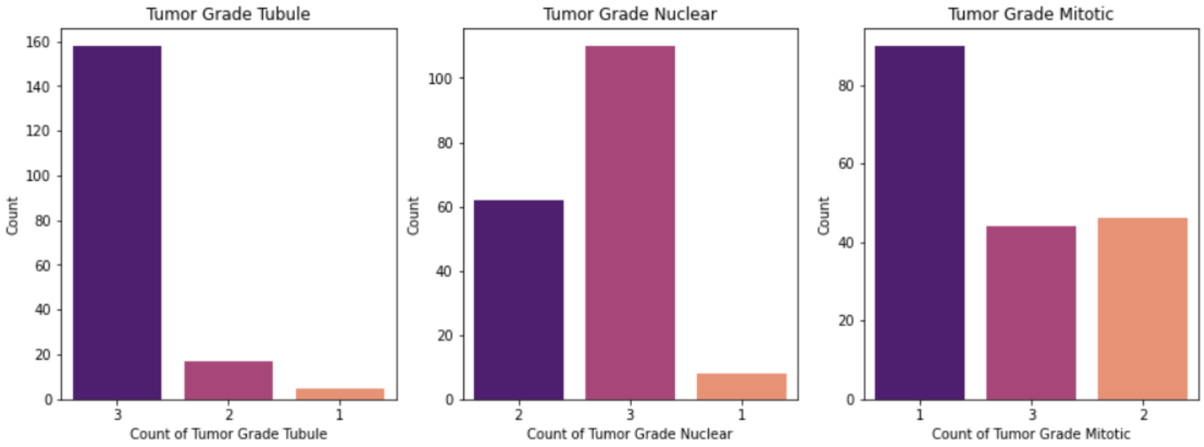
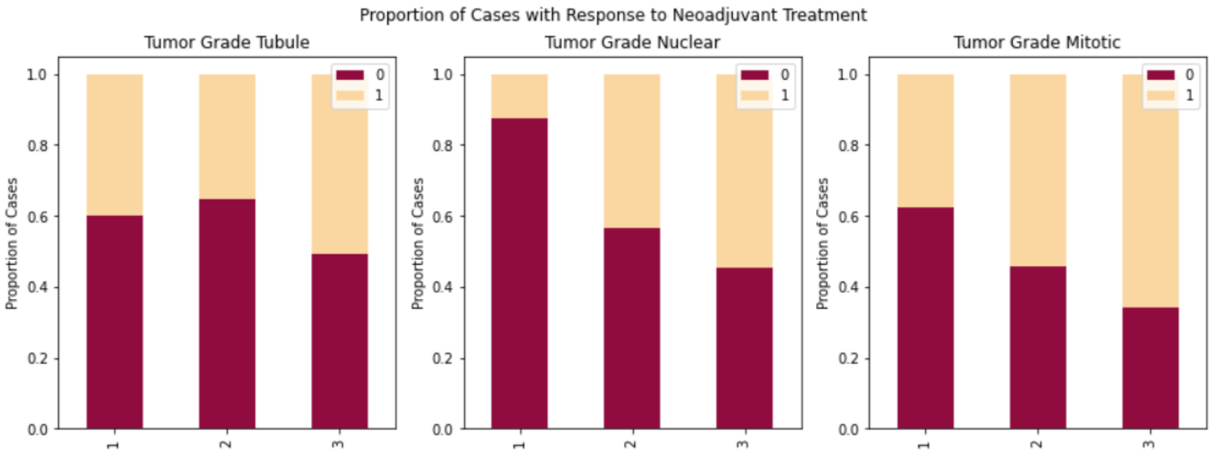


Figure 7. Proportional representation of each grading of tumour type, split by response to treatment. (1 = full or partial response, 0 = no response)



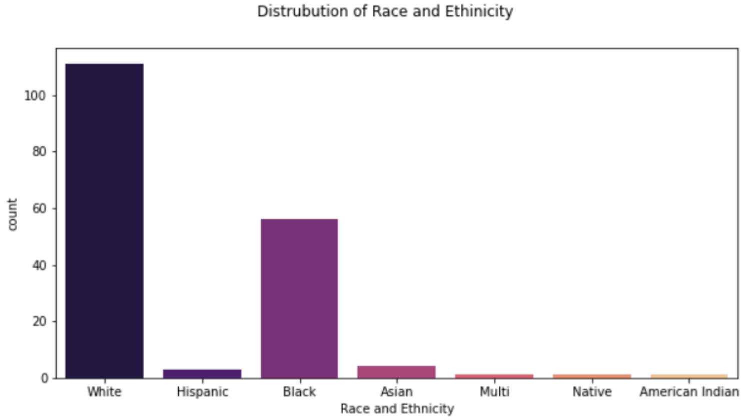
To determine Nottingham grade each of the category grades are combined: a score of 3, 4, or 5 = Grade 1, 6 or 7 = Grade 2, and 8 or 9 = Grade 3. Since we are keeping each individual category Nottingham grade is unnecessary information so was removed. Histologic type refers to the different types of cancer. In our data we have nine different categories: Ductal carcinoma in situ (DCIS), ductal, lobular, metaplastic, Lobular carcinoma in situ (LCIS), tubular, mixed, micropapillary, colloid, mucinous, medullary. This category is deeply imbalanced with 90% of the data being DCIS; this in combination with the high number of NA values 'Histological type' category was also removed.

Demographic factors - Ethnicity

Race and Ethnicity has a very important role in breast cancer, there have been many studies over the last few decades pertaining to certain groups of individuals being more susceptible to cancer. Ashkenazi Jewish women have a 1 in 40 chance of having a mutation in the BRACA gene which increases risk of cancer (BCUK, 2022).

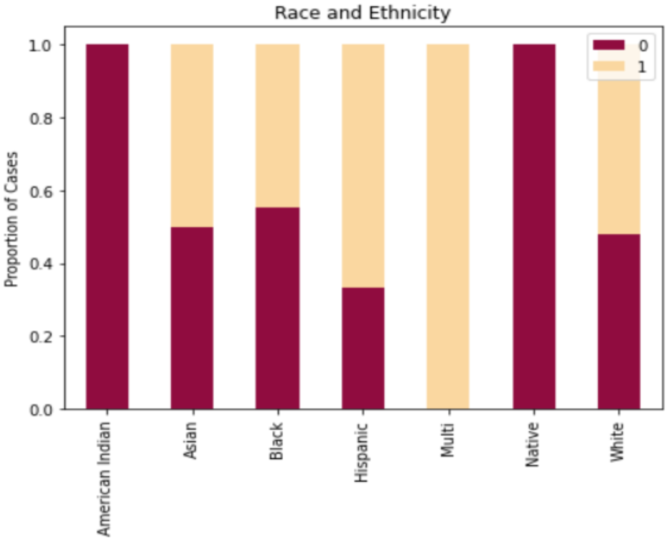
Unfortunately, with medical data there often an unconscious bias; a lot of data curated and collected over the years has been found to have an inherent imbalance favouring those who are white. There have been quite a few examples in machine learning where engineers have not considered or ignored the implication of race on the outcome of their models. For example, a research paper by Obermayer et. al (2019) analyses a widely used algorithm which exhibits racial bias. This bias occurs because the machine learning model utilised health care costs in addition to illness. Unequal access to care in America meant that the lowest socio-economic groups who couldn't afford healthcare and therefore either self-discharged or avoided treatment had a higher proportion of black patients; this group therefore had predicted shorter stay times. The bias was so apparent that when the issue was addressed an increase in the percentage of black patients receiving additional help from 17.7% to 46.5% was observed. In an article by Smeigal et al. (2006) death rates in African American women remain 37% higher than in Whites, despite lower incidence rates. This could also reflect the socio-economic situation in America and reflected inequality in the treatment processes of patients or the possibility that different races can have different reactions to treatment types.

Figure 8. Distribution of each category of race and ethnicity, demonstrating a clear imbalance to white patients



As we can see from the figure above our dataset shows a clear imbalance: we only have one instance of multi, Native or Hawaiian. Therefore, when we compare the number of cases for each, it causes significant of problems. It could be argued that although there has been a lot of research in relation to cancer susceptibility and prognosis with race and ethnicity, there might not be a significant difference in response to treatment. There is no way for us to tell with the limited data that we have. Furthermore, if we are really examining the intersection of race and ethnicity, we need more information specifically about Ashkenazi Jewish heritage or other family history. Therefore, although it is very important to consider these elements, for this model we will have to remove them. Figure 9 illustrates the how this imbalance in the data can provide false equivalence (e.g., 100% of American Indians don't respond to treatment).

Figure 9. Proportional distribution of each category of race and ethnicity, demonstrating how imbalanced data can provide misrepresentation



After data cleaning, removing NA values and one-hot-encoding categorical data, the table below clearly shows the distribution of each category. Note some of the columns in this table are removed after the feature selection process which expands the final dataset to 185 patients as some NA rows no longer need to be removed. In retrospect to keep more instances, the -1 values for staging (Nodes) should not have been dropped, if there was a change in the tumour size that automatically means the response to treatment would be classified as a 1 without the need to know the outcome for change in the node.

Neoadjuvant Clinical Data (176 Patients)		
Column name	**Categories**	**Instances**
Date of Birth (Days)	Min 10183, Max 26902	
	Average 18310, std +- 4082	
Menopause (at diagnosis)	Pre-Menopause	97
	Post-Menopause	79
Race and Ethnicity	American Indian	1
	Asian	4
	Black	56
	Hispanic	3
	Multi	1
	Native	1
	White	66
Contralateral Breast Involvement	No	107
	Yes	33
Lymphadenopathy or Suspicious Nodes	No	69
	Yes	107
ER		123
PR		98
HER2		140

Mol Subtype	Luminal A	97
	HER2	9
	Luminal B	27
	Triple Negative	43
Staging (Tumour Size)	I	36
	II	92
	III	38
	IV	10
Staging (Nodes)	I	82
	II	58
	III	20
	IV	16
Pathologic response to Neoadjuvant therapy: Pathologic stage (T) following neoadjuvant therapy	Complete Response	19
	I	86
	II	52
	III	18
	IV	1
Pathologic response to Neoadjuvant therapy: Pathologic stage (N) following neoadjuvant therapy	Complete Response	82
	I	57
	II	21

	III	16
Response to Treatment	Complete or Partial Response	88
	No Response (Tumour advancement)	88

Feature Selection Using Explainable AI - SHapley Additive exPlanation (SHAP)

To be able to create the most effective machine learning model, we can also apply feature selection techniques. Using SHapley Additive exPlanation (SHAP) we can see which categories may be adding unnecessary complexity to our models and how each variable positively or negatively affects the predicted results. This aids us in our decision to select the most influential parameters.

Figure 10. Mean SHAP values

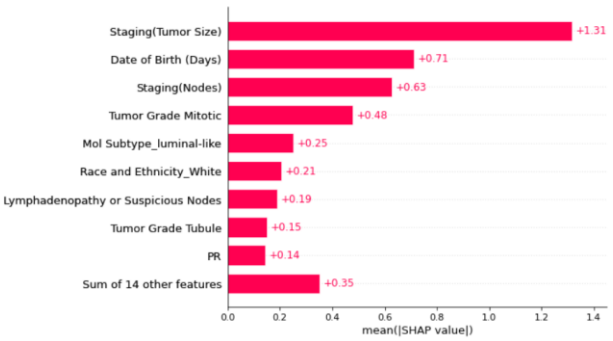


Figure 11. Max SHAP values

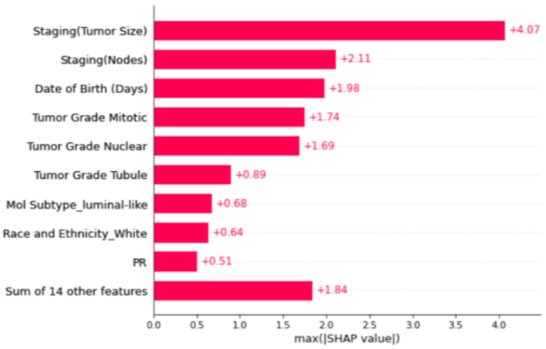


Figure 12. Waterfall SHAP Explainer

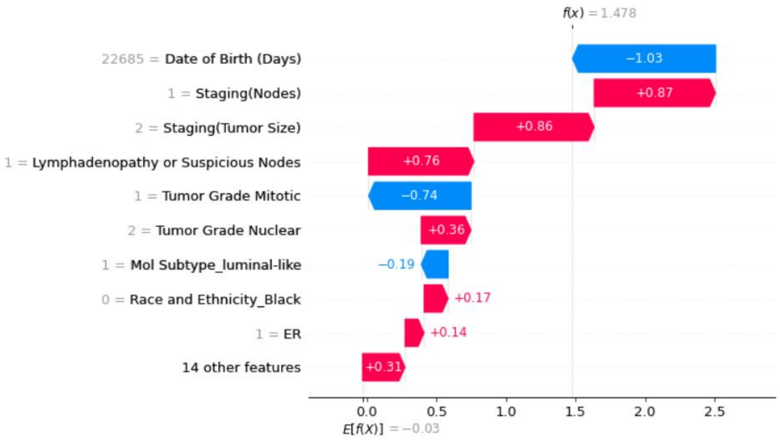
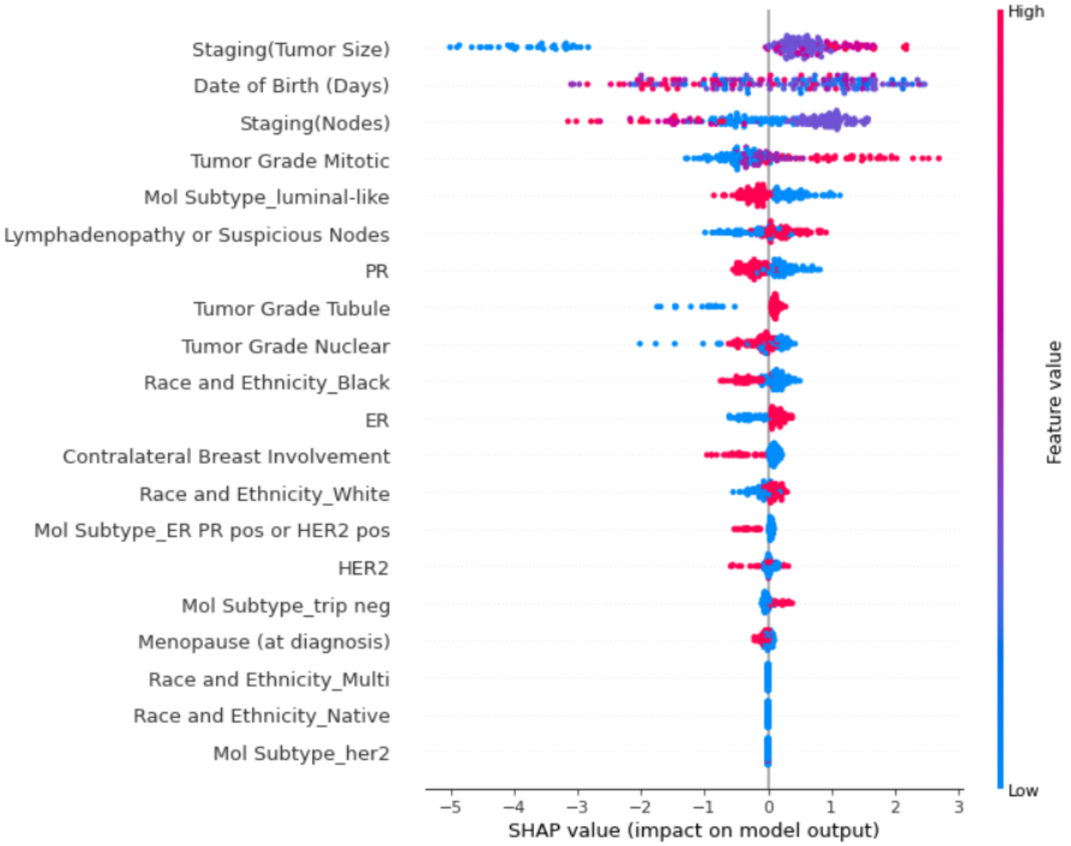


Figure 13. SHAP Tree Explainer



As shown previously by our univariate analysis we can see that 'Staging(Tumor Size)' has a large impact on our model's prediction as does Mitotic tumour grading. Although Date of Birth has a considerable impact on the model's prediction, we can see from figure 13 that there is a mix of positive and negative impacts throughout the range of ages. Through experimentation we tried eliminating this feature to see if this distribution would affect the results, but found keeping it achieved better accuracy.

Nine features were selected after analysis and experimentation: 'Staging (Tumor Size)', 'Date of Birth (Days)', 'Staging(Nodes)' , 'Tumor Grade Mitotic', 'Mol Subtype_Luminal_A', 'Lymphadenopathy or Suspicious Nodes', 'Tumor Grade Tubule', 'PR' and 'Tumor Grade Nuclear'.

Model Architecture

Our data is then split into 70:30 train and validation data, due to the limited size of our dataset we have not created a test set as we need as much information in the training and validation data as possible. A standard scaler is applied to normalise our data, where the standard score of a sample (x) is calculated by:

$$z = \frac{x - u}{s}$$

Where u is the mean of the training set and s is the standard deviation of the training set.

We have used a simple multilayer perceptron with two hidden layers for our clinical data. A difficulty with using deep learning is that models require vast amounts of data; since we have a small dataset in order to prevent overfitting, we have used dropout as a regularisation technique alongside early stopping. In addition, the use of a simple model can also aid in preventing overfitting.

The output of our model is binary, we have used binary cross entropy as our loss function where the loss is the negative average of the log of corrected predicted probabilities. This can be represented by the equation:

$$\log\ loss\ = \frac{1}{N}\sum_{i=1}^{N} -(y_i * log(p_i) - (1 - y_i) * \log(1 - p_i))$$

We have also used Adam optimizer, which uses the squared gradients to scale the learning rate like RMSprop and it takes advantage of momentum by using moving averages of the gradient instead of gradient itself like stochastic gradient descent with momentum. Various learning rates were tested including a slower initial learning rate (0.0001) using an exponential decay scheduler, which had very good training accuracy but was prone to overfitting causing poor validation results. Finally, we have used ReLU activation function which can simply be expressed as:

$$f(x)\ =\ max(0, x)$$

And a Sigmoid output activation function which has a characteristic s curve output, which can be expressed as:

$$S(x)\ =\ \frac{1}{1\ +\ e^{-x}}$$

Where e is Euler's number.

Hyperparameter	**Type**	**Neurons**
	Clinical Model	
Input Layer	Dense	64
Activation function	ReLU	
Hidden Layer	Dense	128

Activation function	X 2	ReLU	64
Dropout		0.2	
Output Layer		Dense	1
Activation function		Sigmoid	
Loss function		Binary Crossentropy	
Optimizer		Adam: Learning rate = 0.01	
Early Stopping Callback		Monitor: validation accuracy, patience :10, best weights restored	
Batch Size		4	
Max Epochs		80	

Hyperparameter Tuning

For further fine tuning, Keras HyperParameter tuner with Bayesian Optimization was also run to produce a different network which included 4 layers with various amounts of dropout in each layer. After cross validation, it was found our original simpler model was more effective, reducing the amount of variance and standard variation, hence our final model uses that architecture. Although keras tuner did achieve a maximum validation accuracy of 86%.

Image Pre-processing - DICOM Extraction

The first challenge of dealing with medical imaging is formatting the data in a form which our models can process. For each patient there are 90-164 slices from MRI in DICOM format with each patient having a variance of dimension in the x and y also. The setting of the MRI scanner can also mean that an individual MRI scan can have different slice thickness and pixel spacing which if not considered could cause variances or lack of accuracy when converted into an array. There are various analyses that simply use CV2/imshow/PIL module to read the DICOM files and then extract a certain number of slices at regular intervals. The method used here instead removes all variances from irregularities from different MRI machines. An array of zeros is created based on the pixel spacing and slice thickness dimensions found in the DICOM file then the image is resized to fit the array using the Pydicom module, eliminating any inconsistencies.

In addition, the Scipy.ndimage zoom module is used to create uniformity in x,y, z dimensions. For each unique patient folder the DICOM images are extracted and resized to three different dimensions 256x256x80 (a common medical imaging dimension), 128x128x60 and 64x64x40. For each resized dataset each patient corresponds to an array containing the image array, corresponding label value (0 or 1) and unique patient identifier. The whole dataset of arrays is then saved as an npy. file, reducing the dataset from 80GB to 8.87GB, 1.75GB and 213MB respectively, which can simply be reloaded in RAM for direct use in the model instead of loaded in situ. Figure 14 and 15, are the visualisations of two patients taken from the converted numpy arrays, 16 slices taken equally across the 80 total slices.

Figure 14. Patient with no response to neoadjuvant treatment

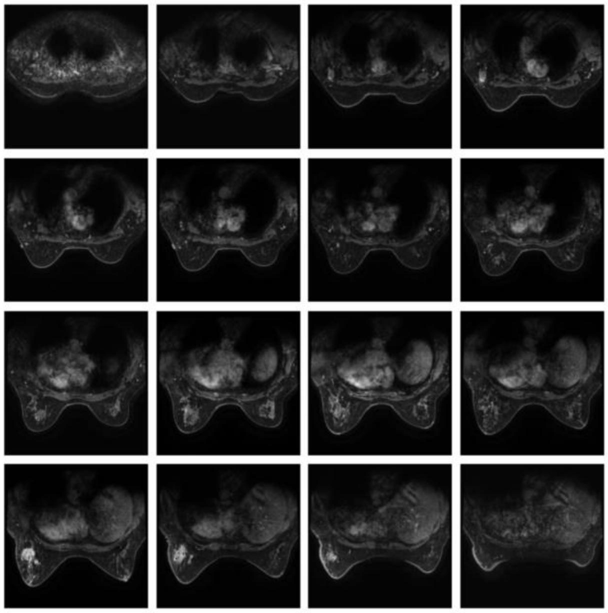
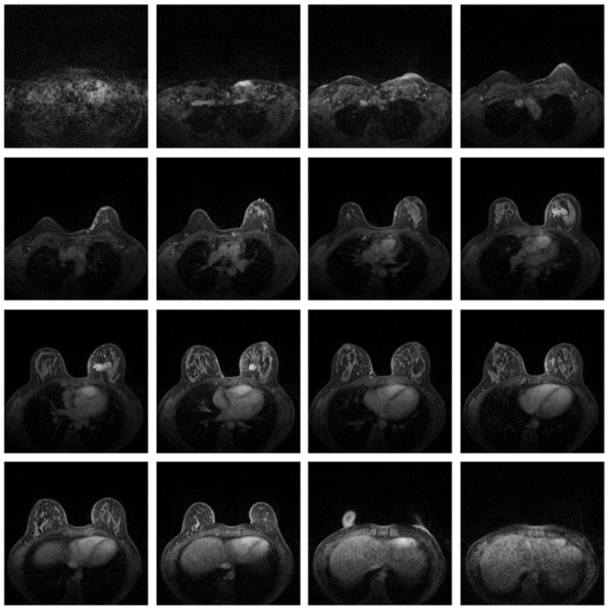


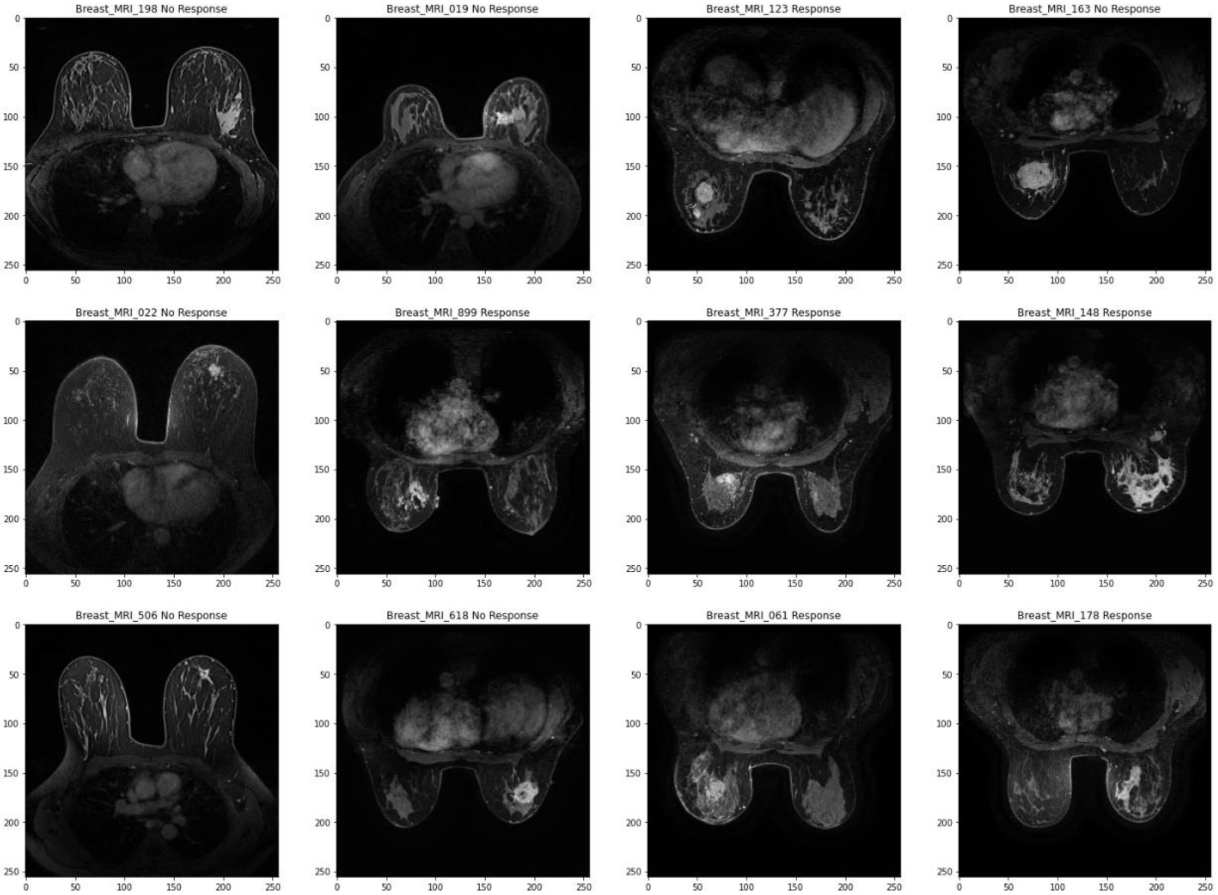
Figure 15. Patient with response to neoadjuvant treatment



At this stage, we can already predict some limitations we might encounter. These images are highly complex and are quite large in size; this means that for such a small dataset our deep learning models are prone to overfitting and might not even be able to capture the relationship between what represents a person who will have a response and who will not. In addition, despite configuring TensorFlow to use GPUs (Graphics Processing Unit) and using mirrored distribution strategy there is a high chance of reaching a memory allocation error (OOM).

In anticipation of this we also extracted a single slice from each patient to create a dataset of 2D images. We achieved this by using the supplementary data from Saha et al. (2018). Included are corresponding annotation boxes which were manually drawn by specialists to segment the largest area of cancer. From this we used the range of slices and took the middle number to choose the corresponding DICOM file slice to extract. Although the middle slice might not necessarily be the largest area of cancer for every patient, it still should show a fair representation of the extent and aggression of the cancer. Figure 16 displays a single slice from a range of patients who have response to treatment and those with no response.

Figure 16 A random selection of patients showing a single slice (determined by annotation boxes) who have and have not has response to neoadjuvant treatment


Breast_MRI_198 No Response
Breast_MRI_019 No Response
Breast_MRI_123 Response
Breast_MRI_163 No Response
Breast_MRI_022 No Response
Breast_MRI_899 Response
Breast_MRI_377 Response
Breast_MRI_148 Response
Breast_MRI_506 No Response
Breast_MRI_618 No Response
Breast_MRI_061 Response
Breast_MRI_178 Response
0
50
100
150
200
250

From a superficial perspective the patients where the tumours(s) are smaller and more compact tend not to have a response, from our clinical analysis this correlates to our finding on how tumour size grading is positively correlated to response to treatment.

Image Model Architecture 2D and 3D

Data is loaded using Tensorflows tf.dataset API, then using an iterator batches of images are prefetched using tf.data.AUTOTUNE; this aids the reduction in computational overload. A pre-processing function is also mapped to each image to standardise the data and apply image augmentation. Image augmentation is a valuable technique used as a regularisation method, although it does not make the dataset bigger it alters the images randomly to create more variety giving the illusion of more data.

Our 2D architecture worked better than our 3D, due to the limited dataset the 3D model either would predict entirely one class or predict randomly. Another problem encountered was memory allocation error, so we had to ensure that the models were not too complex. This was particularly difficult for our 3D model and might explain the poor results from this model, further research is needed to improve our 3D image analysis. To overcome the memory allocation error, some sacrifices had to be made including smaller network architecture in tandem with a small batch size.

Below depicts the final architecture chosen after experimentation and hyper parameter tuning for our 2D and 3D models.

Hyperparameter		**Type**	**Neurons**
2D Image Model - Convolutional filters (3x3)			
Input Layer		Convolutional 2D	32
Activation function		ReLU	
BatchNormalization			
Hidden Layer	X 3	Convolutional 2D	64 256 256
Activation function		ReLU	
BatchNormalization			
Global Average Pooling 2D			
Fully connected layer		Dense	128
Activation function		ReLU	
Output Layer		Dense	1
Activation function		Sigmoid	
Loss function		Binary Crossentropy	

Optimizer	Adam: Learning rate = 0.01
Early Stopping Callback	Monitor: validation AUC, patience :10 , best weights restored
Batch Size	4
Max Epochs	50

Hyperparameter		**Type**	**Neurons**
3D Image Model - Convolutional filters (3x3x3), Pooling filters (2x2x2)			
Input Layer		Convolutional 3D	64
Activation function		ReLU	
MaxPooling3D		Pool size : 2	
BatchNormalization			
Hidden Layer	X 3	Convolutional 3D	64 128 256
Activation function		ReLU	
MaxPooling3D		Pool size : 2	
BatchNormalization			
Global Average Pooling 3D			
Fully connected layer		Dense	512
Activation function		ReLU	

Dropout	0.3	
Output Layer	Dense	1
Activation function	Sigmoid	
Loss function	Binary Crossentropy	
Optimizer	Adam : Exponential decay, initial learning rate = 0.0001	
Early Stopping Callback	Monitor : validation AUC, patience :10 , best weights restored	
Batch Size	4	
Max Epochs	50	

Through experimentation, we found that the removal of the pooling layers in our 2D model improved our accuracy. This is because in the early stage of the networks, discriminative details can be lost due to improper pooling mechanisms. Pooling layers aggregate features from a local region and are included mainly for down sampling feature maps. This can be useful for reducing computational complexity and help the CNN to learn constant features. There are several different types of pooling, Max Pooling and average pooling are the most common forms of pooling, but each come with their own merits and disadvantages. Maxpooling selects the largest value in a region, eliminating unwanted background features. However, as only the maximum value is selected, the pooled representation may capture noisy features. In comparison, averaging reduces the effect of noisy features but has less discriminative power due to the presence of background regions in the pooled representation. The following equations represent the pooling process for max pooling and average pooling respectively. (Manivannan et al. 2022)

$$f_{max}(x) = max\{x_i\}_{i=1}^{N}$$

$$f_{avg}(x) = \frac{1}{N}\sum_{i=1}^{N} |x_i|$$

Transfer Learning

Another method of combating our small data problem is to use transfer learning, a technique of utilising learned features from the weights of pre-trained models. We have used Mobilenet version 2 as it is particularly lightweight compared to other architectures. Despite this there was still an inherent overfitting problem: our training data reached close to 100% accuracy whilst the test data could not surpass 52%. A technique used to combat this was after freezing the base layers we unfroze the batch normalisation layers then added a couple of fully connected layers with dropout, this saw an improvement of 5% in our test data.

Despite the benefits of transfer learning, most architectures have been trained on RGB value images therefore have 3 channels. Our images are grayscale and the process of converting the shape of our arrays to match 3 channels has caused some problems. Below are figure 17 and figure 18, showing how the images look after the pre-processing function is mapped. Both have also had mobilenets pre-processing function applied as per documentation. On the left is the tf.tile method which causes considerable problems and on the right is using the tf.image.greyscale_to_rgb function which works better but we can still see a lot of detail is lost. This may be a reason for why transfer learning is not as effective on our data as we hoped.

Figure 17. visualisation of breast DCE-MRI slice after tf.tile and pre-processing

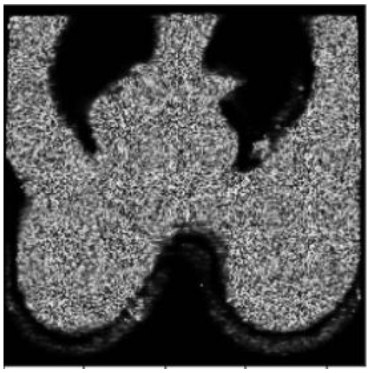
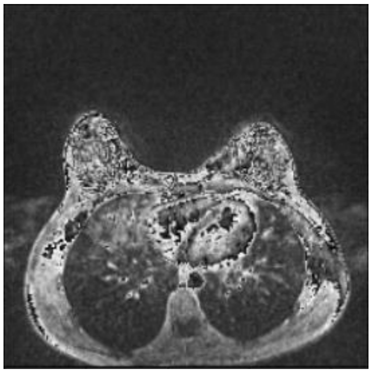


Figure 18. visualisation of breast DCE-MRI slice after tf.image.greyscale_to_rgb and



Late Fusion Methodology

To test our late fusion predictions, we need to create an end-to-end pipeline that will allow two separate models to run and ensure the test set that has both image and clinical data for each patient. The pipeline begins with data cleaning of the clinical data to remove NA values, and this is saved as a new file, a list of patient ID numbers are used to extract the appropriate patient MRI scans. DICOM images are converted into numpy arrays and saved as a npy. file along with patient ID number and target label. The npy. file is converted into a dataframe and merged on the patientID number with the cleaned clinical data to ensure there is no mismatched information. At this point the data can be shuffled, the concatenated data frame is then split into test and train sets. Each of those sets are split into image data and clinical data removing the patient ID number. It is important to make sure that once split, the data is not shuffled as this means each index will not match to an individual patient. After each model is trained the weights are saved. Both model and clinical data predict a value between 0 and 1 from the test set.

At this point there are several approaches that can be used to fuse the predicted results from both models. Max fusion; where if both models predict a value above a threshold then it is classed as a positive result or vice versa depending on the sensitivity or specificity of our model. This also can be manipulated by lowering the threshold for positive classification particularly if false negatives are to be avoided.

Another method is discussed in a paper by Morvant et al. (2021) where they explore PAC-Bayesian Majority Vote for Late Classifier Fusion which aims to minimise the misclassification rate of the Q-weighted majority vote by taking into account the diversity of the voters. MinCq is a quadratic program coming from the machine learning PAC-Bayes theory. They propose an extension of MinCq by adding an order preserving pairwise loss for ranking, helping to improve the Mean Averaged Precision measure. This is a method to potentially explore in further research.

For this paper we have used a form of weighted average fusion; the idea behind using a weighted average is that if our image models predictions have better specificity then this might improve the results from our clinical data in predicting true positives. Our model finalised with an average using 60% clinical model prediction and 40% image model prediction.

Results and Discussions

Through experimentation we tried multiple versions of image sizes; the table below depicts experimentation with best results. As seen below, using late fusion at the decision level we have achieved an accuracy of 82% and an F1 score of 84%.

Binary Classification of Response to Neoadjuvant Treatment							
		Accuracy	**Loss**	**Precision**	**Recall**	**F1**	**AUC**
Full Clinical Data		0.660	0.848	0.622	0.852	0.690	0.66
Feature Selected Clinical Data	Best	**0.755**	**0.513**	**0.733**	**0.815**	**0.772**	**0.824**
	Stratified KFold Cross Validation 4 times						
	Average	0.722	0.556	0.725	0.716	0.720	0.782

		Variance	0.054	0.040	0.051	0.098	0.067	0.052
		Standard Deviation	0.063	0.046	0.059	0.114	0.078	0.060
Image Data	126x126x60		0.536	4.617	0.583	0.467	0.519	0.516
		Stratified KFold Cross Validation 4 times						
	MobileNet V2 256x256	Average	0.562	0.737	0.572	0.652	0.609	0.563
		Variance	0.245	0.082	0.057	0.215	0.090	0.636
		Standard Deviation	0.028	0.951	0.66	0.248	0.361	0.073
	256x256	Best	**0.589**	**0.693**	**0.578**	**0.867**	**0.694**	**0.624**
		Variance	0.058	0.450	0.173	0.193	0.182	0.036
		Standard Deviation	0.067	0.519	0.200	0.223	0.211	0.041
Late Fusion - 2D		Best	**0.821**	**-**	**0.813**	**0.867**	**0.839**	**-**

Where or evaluation metrics are defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

The figures below depicts the results for our clinical data model which can be found in the notebook "Neoadjuvant Treatment Clinical Data Analysis". As we can see from the results, we achieved a promising 0.755% accuracy with an 0.772% F1 score and 0.824% AUC score. In addition the figures below show success in overcoming the overfitting problem quite well. Through stratified K fold cross validation (4 folds) we have also confirmed the robustness of our model. The hyperparameters chosen, perform well in generalising and the results suggest that it would perform well on unseen data.

Figure 19. Visualisation of validation and training model accuracy, loss, and AUC score for clinical model

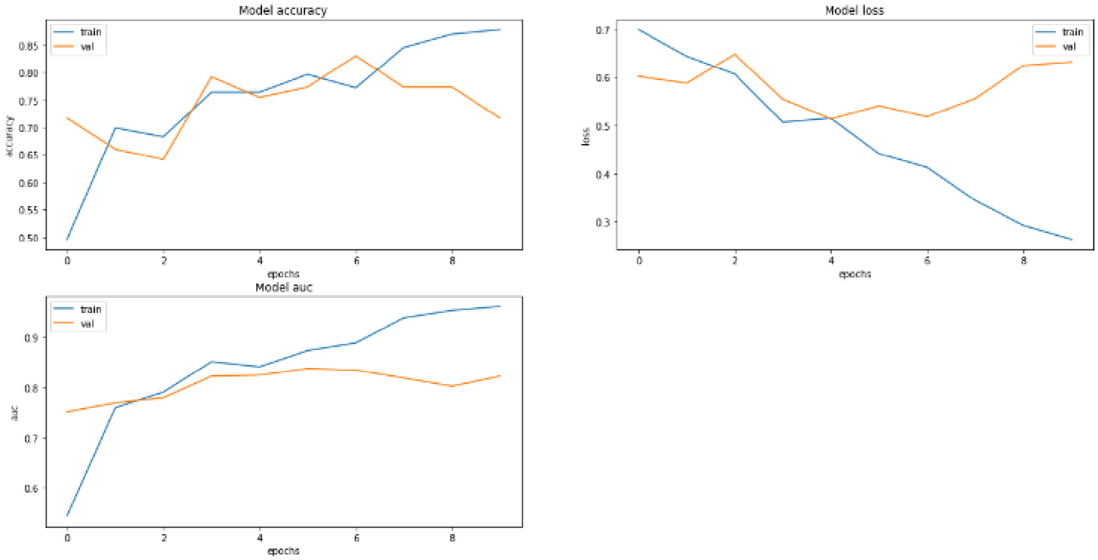
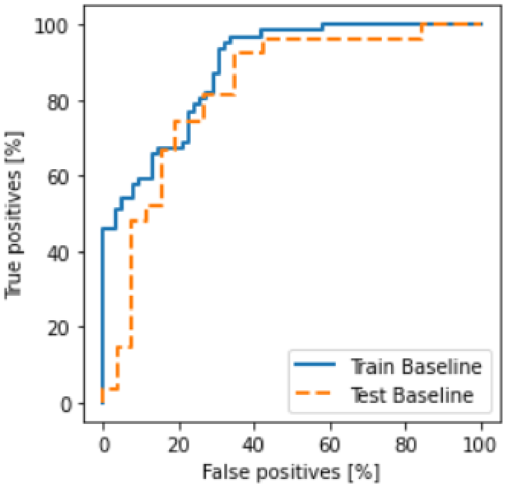
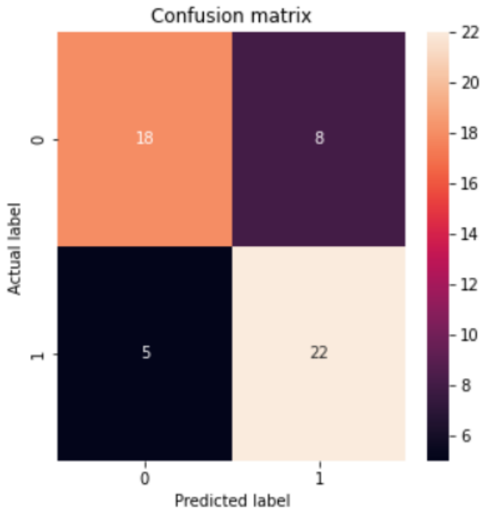


Figure 20. ROC-AUC curve for the clinical model

Figure 21. Confusion matrix of validation data





On the other hand, our image models were not as successful, due to the limitation in data and the restraints on the GPU (less complex model architecture) there is prevalent overfitting. After experimentation the 3D models had no improvement after hyperparameter tuning and therefore it was decided that that route would need further research before implementation in late fusion. Although we had similar problems with our 2D models (as seen in the figures below and the variance and standard deviation in our results) we instead optimised the model by using early stopping callback on the validation area under curve score, which produced a model which had higher sensitivity (recall score) in the aim that this could enhance the clinical model in positive predictions.

Figure 22. Visualisation of validation and training model accuracy, loss, and AUC score for 2D image model

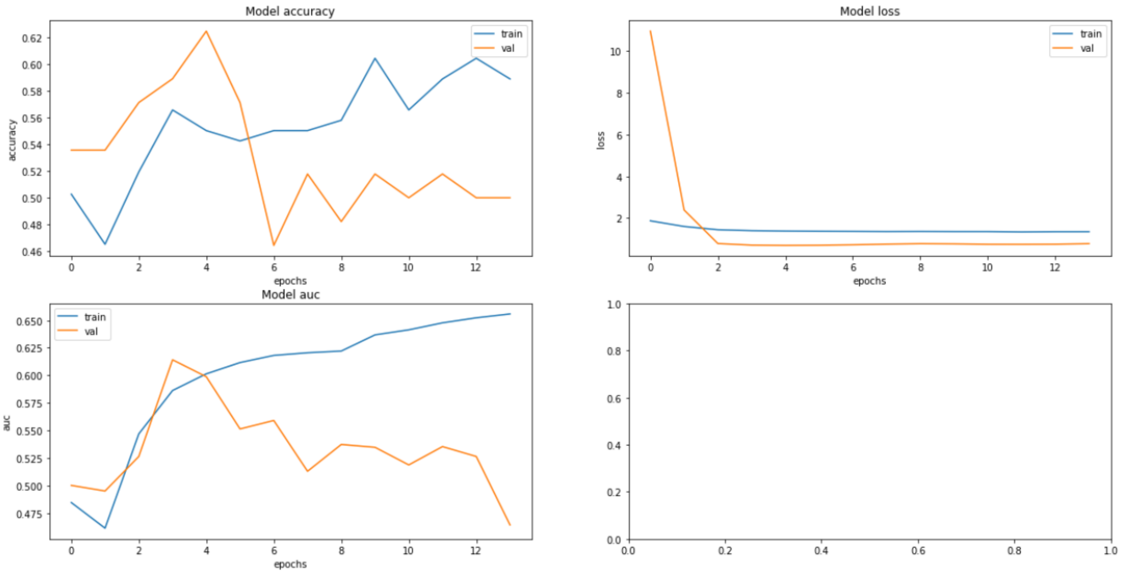


Figure 23. ROC-AUC curve for the 2D image model

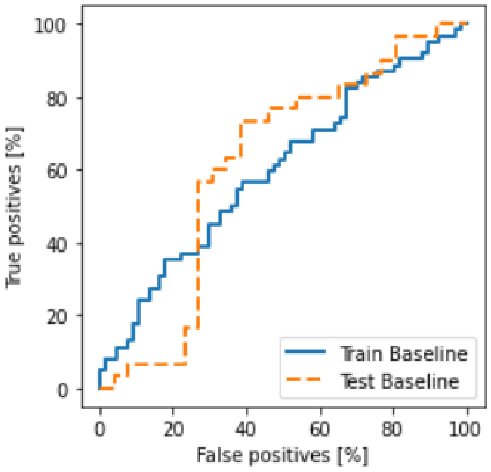
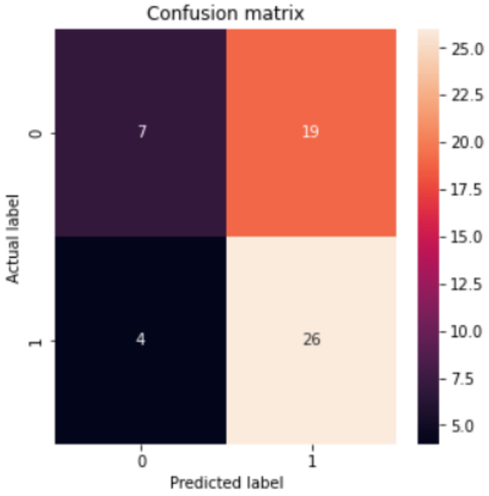


Figure 24. Confusion matrix of validation data



To summarise, although we have achieved an improvement in accuracy through late fusion, due to the limited test sample size there is not a significant change to justify saying that this model is a confident success. With the addition of more data, we would have a much clearer idea on the efficacy of this technique. On the other hand, we can see from our clinical model that there is potential to create an impactful predictive tool for practitioners, to be able to identify the individuals who would not benefit from treatment, which could mean receiving surgery more quickly. In addition, we have successfully fine-tuned our model to reduce the overfitting problem.

In further research we could expand this model to a multi-classification problem, to classify groups of patients dependent on the degree of change in tumour staging. This would highlight those individuals who had advancement in their cancer whilst undertaking neoadjuvant treatment and could be deemed as the highest risk patients.

Supplementary Evidence

Since a huge problem with our experimentation was due to limitations the in data, we also tried the experimentation with a different target variable. "Recurrence Event(s)" refers to if a patient has had cancer return after a period during which the cancer could not be detected. After data cleaning and feature selection processes this allows us to use 834 patients, despite this there is considerable imbalance with less than 10% of cases being positive of recurrence. Recurrence is such a difficult thing for a model to predict, there are many more complexities to why cancer might recur than just assessing an image or biometric markers. Despite this there are still elements found in images and clinical data that might aid in suggesting recurrence, such as high grade of tubules and high grade of nucleic cells. Taking all these factors into consideration we will use recurrence as a target variable and compare how the late fusion model works when we have a much larger volume of data.

Binary Classification of Recurrence of Breast Cancer

		Accuracy	**Loss**	**Precision**	**Recall**	**F1**	**AUC**
Clinical data including treatments		0.617	0.881	0.122	0.500	0.196	0.585
Feature Selected Clinical Data		0.714	1.776	0.125	0.333	0.182	0.540
Image Data	128x128x60	Increased computational power needed					
	64x64x30	0.524	0.752	0.132	0.600	0.216	0.586
	MobileNet V2 256x256	0.667	0.622	0.115	0.375	0.176	0.541
Late Fusion -2D		**0.806**	**-**	**0.162**	**0.25**	**0.199**	**-**

As seen in the table above, all of our prediction's precision and recall scores are fairly low, due to the highly imbalanced nature of the data. Despite this, for our late fusion model we have still shown an increase in accuracy from 71% in the clinical data, 67% in the image data to 80% through late fusion.

Challenges, Limitations and Future work

Although we have achieved positive results showing improvement in accuracy and F1 score for our data, there is still a lot of work needed to prove the efficacy of this technique; limitations in our dataset size has hindered the potential in performance. Despite this, what we can conclude is that the viability of a model that can predict response to neoadjuvant treatment is feasible, even with limited data we had very positive results. The most immediate solution would be to increase our dataset size.

To enhancing our clinical data model, we could also use other features such as biometric markers (other than ER, PR, HER2) or genome expression. In the paper "Predicting the response to neoadjuvant therapy for early-stage breast cancer: tumor-, blood-, and imaging-related biomarkers" (Tan et al. 2018) the authors explore in detail all the various elements which could be measured to aid predicting response to neoadjuvant treatment. For example, elements include some multi-gene assays, such as, Oncotype DX, MammaPrint, and Prosigna. Despite this, the author warns that "although biochemical biomarkers from peripheral blood and/or tumor tissue appear promising, there is still a lack of consensus in practice guidelines to guide the NAT of breast cancer." Other information that could be insightful to our model is family history. We already know that genetic predisposition to cancer is hereditary but there is not much research into whether that information could aid in prediction for the treatment process. Similarly with race and ethnicity, the dataset used for our analysis had a severe imbalance and ultimately was removed due to limitations, but there could also be insightful information there.

Ultimately the most potential for future work is enhancing our image models, whether that is by increasing our dataset, computational power or even using a different image analysis technique. If we are to use more data, there will be a need to use multiple GPUs with distribution strategy. Another option is to use TPUs (Tensor Processing Unit), but it must be noted that a difficulty with using TPU is the lack of compatibility with Tensorflows data generators. In terms of improving the variance in our models results there are also various ensemble learning techniques which could be utilised, which average results by combining several variations of models.

Another method that has had proven success in literature is image segmentation. This would focus our data and reduce background information that might be confusing our models. Or use machine learning to extract features from images such as volume and density, we could then use feature extraction and selection techniques to reduce the complexity of the imputed data. A paper by Virtrosko et al (2021) shows the practical application of this kind of technique, using parameters characterising the change in tumour volume, cellularity (ADC), and vascular characteristics (Ktrans) to predict pathological complete response to NAT which they found to be significantly better than the Response Evaluation Criteria in Solid Tumours (RECIST) criteria. Although an interesting study demonstrating the ability for these sorts of models to be used in a practical setting, this study was only done with 28 women.

Finally, we could use a different transfer learning model. A paper by Alzubaidi et al. (2021) proposes MedNet, a model which is specifically trained on medical images. The proposed technique is based on training Gray-MedNet using three million publicly available greyscale medical images including MRI, CT, X-ray, ultrasound, and PET. This would be particularly useful for our dataset, as the conversion to 3 colour channels in pre-processing had caused some issues.

Conclusions

To conclude, we have successfully achieved 82% accuracy and a F1 score of 84% for the classification of response to neoadjuvant therapy. Despite limitations in the size of our dataset this still shows potential in our proposed method of late fusion, as we have seen an increase from 76% accuracy, 77% F1 score in our clinical model and 59% accuracy, 70% F1 score in our image model to 82% accuracy, 84% F1 score in fusion. It must also be taken into consideration due the limited size of the dataset there is a considerable amount of variation and deviation in our model's predictions, therefore although we have seen a positive result there is still need for development. It was found the use of 2D images were far more successful than the 3D scans, this could be attributed to the limitation in computational processing power or too much information for our model to determine the best weights.

In addition, our proposed method also worked well on a very imbalanced larger dataset using recurrence of cancer as a target variable. Seeing an increase from 71% in the clinical data, 67% in the image data to 80% through late fusion. Prediction recurrence is very difficult and not the most suitable variable for machine learning as there are a lot more external factors that can influence this decision. Despite this our model shows potential for development, the next step would be to overcome a severe imbalance problem. We have also found that there is a positive correlation between tumour size, nuclear grade, and mitotic grade respectively to response to treatment. Triple negative and HER2 positive patients were the most responsive to neoadjuvant treatment.

For future work we suggest after collecting more data, to use MedNet pre-trained weights and more complex model architecture to improve our model's prediction. If this is not successful, then either image segmentation or feature extraction. From this point we can then assess other late fusion techniques such as majority voting or use ensemble learning in conjunction. Overall, our study has found there is potential in improving classification of response to neoadjuvant therapy through late fusion, the extent of how much this can be improved is yet to be evaluated and further work on a larger dataset is needed.

References

DUKE BREAST CANCER DATASET

Saha, A., Harowicz, M.R., Grimm, L.J., Kim, C.E., Ghate, S.V., Walsh, R. and Mazurowski, M.A., (2018) 'A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features'. *British journal of cancer*, 119(4), pp.508-516.

OTHER REFERENCES

Antropova N, Huynh BQ, Giger ML.(2017) 'A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets.' *Med Phys.* 44(10):5162-5171. doi: 10.1002/mp.12453. Epub 2017 Aug 12. PMID: 28681390; PMCID: PMC5646225.

Alzu'bi, A., Najadat, H., Doulat, W. et al.(2021) 'Predicting the recurrence of breast cancer using machine learning algorithms.' *Multimed Tools Appl* 80, 13787–13800. Available at : https://doi.org/10.1007/s11042-020-10448-w

Alzubaidi, L., Santamaría, J., Manoufali, M., Mohammed, B., Fadhel, M.A., Zhang, J., Al-Timemy, A.H., Al-Shamma, O. and Duan, Y. (2021) *MedNet: pre-trained convolutional neural network model for the medical imaging tasks*.Available at : arXiv preprint arXiv:2110.06512.

Breast Cancer UK. (2022) *About Breast Cancer Facts and Figures*. Available at: https://www.breastcanceruk.org.uk/about-breast-cancer/facts-figures-and-qas/facts-and-figures/?gclid=Cj0KCQjwl92XBhC7ARIsAHLl9an5z5V7RT78TOK7UFYw0Utjwpaa8btZClhStwgfhPIPhua2UY7Cf5AaAoAwEALw_wcB. (Accessed : 28/07/2022).

Chen S, Shu Z, Li Y, Chen B, Tang L, Mo W, Shao G, Shao F. (2020) 'Machine Learning-Based Radiomics Nomogram Using Magnetic Resonance Images for Prediction of Neoadjuvant Chemotherapy Efficacy in Breast Cancer Patients.' *Front Oncol*. Available at : doi: 10.3389/fonc.2020.01410. PMID: 32923392; PMCID: PMC7456979.

Ha R, Chin C, Karcich J, Liu MZ, Chang P, Mutasa S, Pascual Van Sant E, Wynn RT, Connolly E, Jambawalikar S.(2019) 'Prior to Initiation of Chemotherapy, Can We Predict Breast Tumor Response? Deep Learning Convolutional Neural Networks Approach Using a Breast MRI Tumor Dataset.' *J Digit Imaging*. 32(5):693-701. Available at : doi: 10.1007/s10278-018-0144-1. PMID: 30361936; PMCID: PMC6737125.

Huang, SC., Pareek, A., Seyyedi, S. et al. (2020) 'Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines'. *npj Digit. Med*. 3, 136. Available at : https://doi.org/10.1038/s41746-020-00341-z

Ismael, S.A.A., Mohammed, A. and Hefny, H. (2020). ‘An enhanced deep learning approach for brain cancer MRI images classification using residual networks.’ *Artificial intelligence in medicine*, 102, p.101779.

Leibig, C., Brehmer, M., Bunk, S., Byng, D., Pinker, K. and Umutlu, L. (2022). ‘Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis.’ *The Lancet Digital Health*, 4(7), pp.e507-e519.

Morvant, E., Habrard, A. and Ayache, S. (2021) ‘PAC-Bayesian majority vote for late classifier fusio’n. *arXiv.* Available at : preprint arXiv:1207.1019.

Nirthika, R., Manivannan, S., Ramanan, A. (2022) ‘Pooling in convolutional neural networks for medical image analysis: a survey and an empirical study.’ *Neural Comput & Applic* 34, 5321–5347. Available at : https://doi.org/10.1007/s00521-022-06953-8

Nuffield Trust. (2021) *Cancer Survival Rates*. Available at: https://www.nuffieldtrust.org.uk/resource/cancer-survival-rates?gclid=Cj0KCQjwof6WBhD4ARIsAOi65ajy97P-RZFSNPBc65OAL2stneyPt1lfcKhYCeGyYTTngcriYHiCyAIaAv_GEALw_wcB#background (Accessed: 12/07/2022).

Obermeyer Z, Powers B, Vogeli C, Mullainathan S.(2019) ‘Dissecting racial bias in an algorithm used to manage the health of populations’. *Science.* 366(6464):447-453. Available at : doi: 10.1126/science.aax2342. PMID: 31649194.

Onitilo AA, Engel JM, Greenlee RT, Mukesh BN. (2009) ‘Breast cancer subtypes based on ER/PR and Her2 expression: comparison of clinicopathologic features and survival’. *Clin Med Res*. 7(1-2):4-13. Available at : doi: 10.3121/cmr.2009.825. PMID: 19574486; PMCID: PMC2705275.

Papanastasopoulos, Z., Samala, R.K., Chan, H.P., Hadjiiski, L., Paramagul, C., Helvie, M.A. and Neal, C.H. (2020) ‘Explainable AI for medical imaging: deep-learning CNN ensemble for classification of estrogen receptor status from breast MRI.’ *In Medical imaging 2020: Computer-aided diagnosis* (Vol. 11314, pp. 228-235). SPIE.

Pandeya, Y.R., Lee, J. (2021) ‘Deep learning-based late fusion of multimodal information for emotion classification of music video’. *Multimed Tools App*l 80, 2887–2905. Available at :https://doi.org/10.1007/s11042-020-08836-3

Prakash S. S. and Visakha K. (2020) "Breast Cancer Malignancy Prediction Using Deep Learning Neural Networks," *Second International Conference on Inventive Research in Computing Applications (ICIRCA*), 2020, pp. 88-92, Available at : doi: 10.1109/ICIRCA48905.2020.9183378.

Smigal, C., Jemal, A., Ward, E., Cokkinides, V., Smith, R., Howe, H.L. and Thun, M. (2006), 'Trends in Breast Cancer by Race and Ethnicity: Update 2006.' *CA: A Cancer Journal for Clinicians*, 56: 168-183. Available at : https://doi.org/10.3322/canjclin.56.3.168

Tan W, Yang M, Yang H, Zhou F, Shen W. (2018) 'Predicting the response to neoadjuvant therapy for early-stage breast cancer: tumor-, blood-, and imaging-related biomarkers. Cancer' *Manag Res*. 10:4333-4347. Available at : doi: 10.2147/CMAR.S174435. PMID: 30349367; PMCID: PMC6188192.

Vaidya J S, Massarut S, Vaidya H J, Alexander E C, Richards T, Caris J A et al. (2018) 'Rethinking neoadjuvant chemotherapy for breast cancer' *BMJ* 360 :j5913 Available at : doi:10.1136/bmj.j5913

Vale-Silva LA, Rohr K.(2021) 'Long-term cancer survival prediction using multimodal deep learning.' *Sci Rep*. 11(1):13505. Available at : doi: 10.1038/s41598-021-92799-4. PMID: 34188098; PMCID: PMC8242026.

Verburg, E., van Gils, C.H., van der Velden, B.H., Bakker, M.F., Pijnappel, R.M., Veldhuis, W.B. and Gilhuijs, K.G. (2022). 'Deep learning for automated triaging of 4581 breast MRI examinations from the DENSE trial.' *Radiology*, 302(1), pp.29-36.

Virostko, J., Sorace, A.G., Slavkova, K.P. et al. (2021) 'Quantitative multiparametric MRI predicts response to neoadjuvant therapy in the community setting.' *Breast Cancer Res* 23, 110. Available at : https://doi.org/10.1186/s13058-021-01489-6

Xie J, Liu R, Luttrell J 4th, Zhang C. (2019) 'Deep Learning Based Analysis of Histopathological Images of Breast Cancer.' *Front Gene*t. 10:80. Available at : doi: 10.3389/fgene.2019.00080. PMID: 30838023; PMCID: PMC6390493.

Xie, J. and Zhu, M. (2019) 'Handcrafted features and late fusion with deep learning for bird sound classification.' *Ecological Informatics*, 52, pp.74-81.

Yala, A., Lehman, C., Schuster, T., Portnoi, T. and Barzilay, R. (2019). 'A deep learning mammography-based model for improved breast cancer risk prediction.' *Radiology*, 292(1), pp.60-66.

Zhu, Z., Albadawy, E., Saha, A., Zhang, J., Harowicz, M.R. and Mazurowski, M.A. (2019). 'Deep learning for identifying radiogenomic associations in breast cancer'. *Computers in biology and medicine*, 109, pp.85-90.