Practical Assignment Report 2

Introduction

In the UK all women between the ages of 50 and 70 years of age are invited to be screened as part of the NHS National Breast Cancer Screening Programme (Massat et al., 2016). Women who are suspected to have breast cancer will have their breast tissue biopsied (i.e. sampled), have their tissue *histopathologically* diagnosed-this means that the cells will be analysed under a microscope- and have a cancer diagnosed or excluded (Dillon et al., 2006). Under the microscope cells will be examined by a professional for any signs of cancer; a cancerous cell may for example become more dense (and thus darker and more purple on *Haematoxylin and Eosinophil* staining) as it proliferates (Howard et al., 2021). The histopathological diagnosis is time consuming, requires a high degree of skill and the consequences of missing a cancerous processes are stressful both for the patient and healthcare resources spent in supporting a misdiagnosis (Massat et al., 2016).

The Lancet published a policy paper (Taylor-Phillips et al., 2022) recommending a large-scale prospective randomised controlled trial to observe the effects of implementing Artificial Intelligence (AI) in the National Breast Cancer Screening Programme. Building on this, we have decided to embark on a project wherein an AI model will classify histological (i.e., cellular) slides of breast tissue biopsies, depending on the presence of Intraductal Carcinoma (IDC), a variant of breast cancer (Howard et al., 2021). We hope the automation of this tedious, time consuming and highly skilled task could bring the benefits of faster, more accurate and cheaper diagnoses.

It must be noted that as with any product being with healthcare we must ensure that it is *safe* and more effective than the alternative methods. Some studies have estimated an accuracy rate of 65% (Dillon et al., 2006) of current methods of manually inspecting breast tissue. A systematic review (Co et al., 2023) of AI's capability in detecting IDC has demonstrated high accuracy (83.78%), sensitivity (83.88%), and specificity (85.49%) in AI-assisted histopathological analysis, also noting that Convolutional Neural Networks (CNN) are the most commonly employed Machine Learning (ML) models. In contrast to the trend of developing CNNs for this task, we aim to identify a model with sufficient higher explanatory power, aligning with calls for AI transparency in the healthcare sector (Jung et al., 2023).

Loading and Inspecting our Data

Having decided we would like to procure data of histological slides of Breast ICD, we found the following data set Mooney (2018). Whilst we were attracted with having a substantial (5000 images), high quality dataset. We were disappointed with the lack of information on how the dataset was curated. Another disadvantage is that the dataset was also not well peer reviewed.

We then moved into working in a Google Collab notebook which we have submitted alongside this report. We then performed some basic inspections of our data. The output of cell 4 shows that we have divided our dataset into a training set of 4,437 images and a test set of 1,110 images, we thus have a well-balanced split crucial for effective model training and evaluation. With over 5,500 images in total, the dataset is sufficiently large to support robust model training and to test for generalisability. Uniformity across images, with each having dimensions of 50x50 pixels and RGB colour channels, streamlines the preprocessing for machine learning application. The dataset's binary classification task, aimed at identifying the presence (1) or absence (0) of IDC, informs our strategic choice of models and evaluation metrics. The labels are in the form of one-hot encoded binary vector with the binary values of 0's or 1's. The use of a fixed random state guarantees reproducibility, setting a strong foundation for an Exploratory Data Analysis and predictive modelling with a clear path towards achieving precise diagnostic outcomes.

On the right we have Figure 1. which is a cropped image taken from the output of Cell 5 of our notebook. It is a basic visualisation of the images which we have identified as either being IDC negative or IDC positive. We can see some basic morphological differences between the two classes in line with what we would expect for the two different classes of cells. For instance, IDC negative images appear to exhibit a more ordered structure with a relative lighter density (reflected by the light pinker stain) whilst conversely IDC positive images appear to exhibit more disorder and more density reflected on their more darker, purple staining.

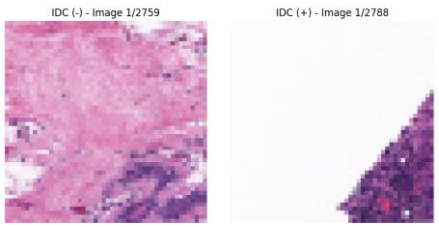

Figure 1. A cropped image of our 'Basic Visualisation of Data' ouput showing images from our IDC positive and IDC negative classes

Preprocessing

In the preprocessing phase of our analysis, two pivotal steps were undertaken to prepare the breast cancer histology images for subsequent machine learning models.

Initially in Cell 7, the images were flattened, transforming them from their original two-dimensional or three-dimensional matrix form into one-dimensional vectors. This transformation is essential for the application of dimensionality reduction techniques, which require data in a format of samples by features to effectively identify and extract significant patterns. Subsequently in Cell 8, we employed *StandardScaler* (a standardisation method from sklearn) to standardise the features by removing the mean and scaling to unit variance, a step that harmonises the scale of features thereby preventing dominance of particular features due to their magnitude. Standardisation can be especially beneficial if your data contains features with high variances or if the features are not bounded (like pixel intensity values). It can also be preferable if you're using algorithms that are based on measures of how far apart data points are, like SVMs, since standardisation keeps outliers in check and preserves their effect.

This not only facilitates a faster convergence of optimisation algorithms but also enhances the accuracy of models that are sensitive to the scale of input features.

Additionally, we refrained from employing image rotation as a preprocessing step, recognising that the cellular structures within the histology images naturally exhibit a wide range of orientations.

Exploratory Data Analysis

Following the preprocessing of our dataset, we embarked on an Exploratory Data Analysis (EDA) to extract insights pivotal for guiding the development of our ML models. Our analysis commenced with a fundamental statistical summary, revealing a median pixel value of 190.00 closely aligned with the mean of 185.03. This proximity indicates a balanced distribution of pixel intensities, suggesting an equilibrium between the IDC positive and negative classes, which are typically characterised by lighter pink and darker purple staining, respectively.

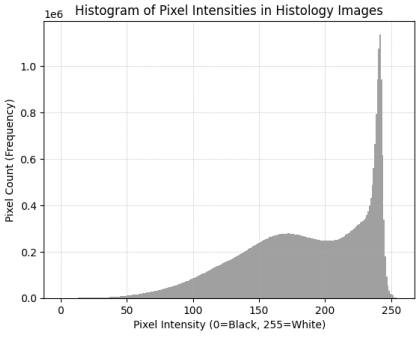

Figure 2. Histogram of Pixel Intensities in Histology Images Indicating Variations Between IDC Positive and IDC Negative Samples

Subsequently, we examined the pixel intensity distribution in greater detail. Figure 2 presents this visualisation, disclosing a bimodal distribution that aligns with our preceding assumptions: IDC negative samples correspond to lower pixel intensities, whilst IDC positive samples are inclined towards higher intensities. The observed bimodal distribution and the balanced mean and median values suggest that preprocessing was effective and that the dataset is well-suited for developing diagnostic models without further need for balancing or anomaly removal.

Statistical characteristics of our pixel distribution—particularly the close alignment of the mean and median, a moderate standard deviation of 47.25, and a slightly negative kurtosis of -0.45—indicate a fairly standard distribution, devoid of pronounced anomalies that could necessitate further preprocessing or impact our model selection strategy.

We then proceeded to employ Principal Component Analysis (PCA), a linear dimensionality reduction technique. Figure 3, illustrating the PCA output, unveiled a significant mingling of the IDC diagnostic classes. The observed overlap signals the potential challenge for linear discriminative models to effectively classify these cases.

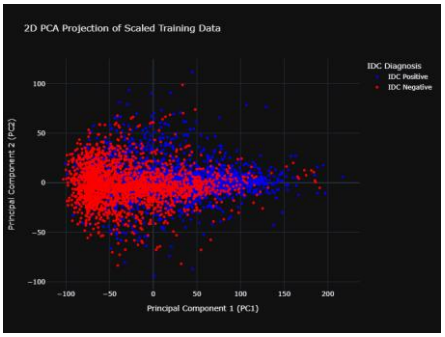

Figure 3. A 2D scatter plot of the first two principal components resulting from PCA performed on the histology images' training set after scaling.

Given these findings, we turned to Uniform Manifold Approximation and Projection (UMAP), an advanced, non-linear dimensionality reduction approach. UMAP maintains the local structure, meaning that points that are close to each other in the high-dimensional space remain close in the reduced space. Contrary to PCA, In the UMAP visualisation, there seems to be more distinct clustering of the IDC positive and negative classes, indicating that UMAP is better suited for datasets where the relationship between data points is complex and not linear., as depicted in Figure 4. It revealed subtle yet distinct groupings (however still not perfect) within our dataset, which a sophisticated classifier could cope with.

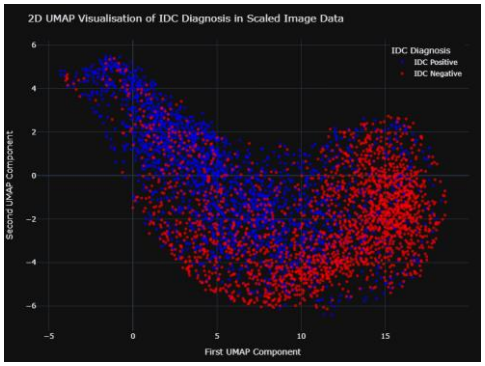

Figure 4. A 2D visualisation achieved through UMAP, which serves to condense the high-dimensional data of histology images into two meaningful components that capture the essence of the scaled histology dataset's structure.

It is also worth to explain hyperparameters used in UMAP. Setting n_components=2 means that UMAP will project the data into a two-dimensional space, which is useful for visualisation purposes as it can be easily plotted. This controls the size of the local neighbourhood used to construct the manifold approximation. Setting n_neighbors=20 suggests a moderate approach, considering both local and more global aspects of the data. This parameter controls how tightly UMAP is allowed to pack points together, which can affect the clarity of the separation between clusters. It essentially controls the trade-off between preserving the local versus global structure of the data. With min_dist=1.0, UMAP allows points to be packed together more closely, potentially leading to denser clusters in the reduced space.

PCA provided a preliminary view of our data's variance, but it was UMAP that delivered the detailed insights guiding our choice of sophisticated, likely non-linear models suitable for the complex task of IDC diagnosis.

Model Selection and Implementation

In choosing suitable models, we have been mindful to strike a balance between the complexity required to navigate intricate data relationships and the need for transparency and interpretability—prerequisites in healthcare applications.

The linear models, constrained by the assumption of linearly separable data, were deemed unsuitable following our UMAP analysis. Consequently, we shifted our focus to non-linear algorithms, specifically Random Forest, SVM with a non-linear kernel and XGBoost. These were chosen for their robust performance and their interpretability features, which are crucial for clinical decision-making.

The Random Forest model, an ensemble of decision trees, is renowned for its high accuracy and feature ranking abilities, providing insights into the variables most indicative of IDC. Our implementation of Random Forest involves 100 estimators and a maximum depth of 5, intending to maintain model simplicity for interpretability while ensuring sufficient complexity for accurate classification.

XGBoost, on the other hand, offers a high level of precision and a suite of interpretability tools, such as SHAP values and feature importance scores. Our XGBoost model utilises 300 estimators and a nuanced combination of hyperparameters, including a learning rate of 0.05 and gamma of 0.5, to balance the bias-variance trade-off. The hyperparameter selections were made to mitigate the risk of overfitting, and to enhance the models' generalisability, ensuring reliable performance across diverse patient datasets.

We also incorporate SVM with a non-linear kernel, which excels in handling complex, high-dimensional datasets. The RBF kernel enables the SVM to establish a non-linear decision boundary, adeptly separating the IDC classes in a manner linear classifiers cannot achieve.

Our evaluation strategy includes a classification report and a confusion matrix for each model, providing an overview of their performance in terms of precision, recall, and F1-score. Furthermore, we will conduct Receiver Operating Characteristic (ROC) curve analysis and compute the Area Under the Curve (AUC) for the model that we will select out of the three.

In summary, our model selection is deeply rooted in the insights derived from the UMAP visualisation and a commitment to the ethical deployment of AI in medical diagnostics.

Evaluation of Models

The SVM with an RBF Kernel displayed a laudable accuracy of approximately 77.75%. This model's precision and recall for both classes were well balanced, indicating a consistent classification ability. The confusion matrix produced (Figure 5) showed a solid true positive rate; however, the presence of false positives and negatives highlighted areas for potential improvement. The interpretability inherent in SVM and the robust performance make it a strong contender for clinical application.

The Random Forest model's accuracy stood at approximately 75.86%, demonstrating a balanced capability to classify IDC. Despite this, the confusion matrix produced (Figure 6) revealed room for reducing both false positives and false negatives to improve clinical utility. Its feature importance metrics provide insightful interpretability, crucial for understanding and trusting AI-assisted diagnostic decisions.

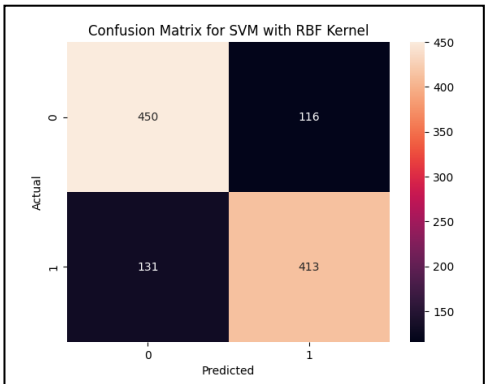

Figure 5: Confusion Matrix for SVM Classifier with RBF Kernel on IDC Classification

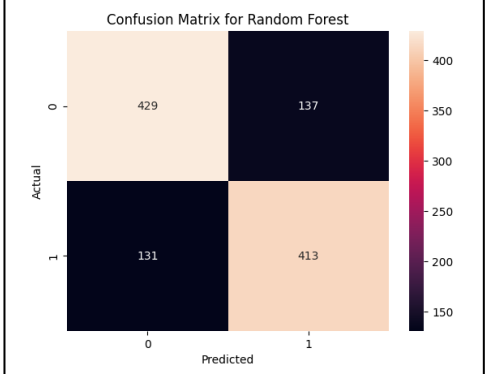

Figure 6: Confusion Matrix for SVM Classifier with RBF Kernel on IDC Classification

XGBoost demonstrated superior accuracy at approximately 78.20%, with a commendable false negative rate that is particularly significant in healthcare where the cost of missed diagnoses is high. The model's precision, recall, and f1-scores show its efficacy in correctly identifying IDC cases. This, combined with its interpretability, underscores its potential for clinical use, where reducing false negatives is critical for patient outcomes. XGBoost's balance of precision and recall, along with its interpretability tools (such as SHAP values), make it highly suitable for clinical use. The model's ability to minimize false negatives is particularly crucial, ensuring patient outcomes are not compromised by undetected cases.

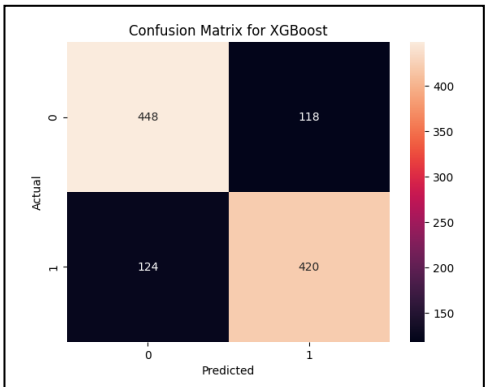

Figure 7: Confusion Matrix for XGBoost Model on IDC Classification

Considering the nuanced requirements of our problem statement—the necessity for high accuracy, transparency, and clinical applicability—XGBoost emerges as the model of choice. It not only fulfills the technical criteria of precision and reliability but also aligns with the ethical imperatives of healthcare AI, offering clarity and insights into its decision-making process. This decision is made with recognition of the need for AI tools that can be seamlessly integrated into existing clinical workflows, enhancing the screening programme's efficacy while maintaining patient trust. With XGBoost selected, we will proceed to further validate its capabilities through Receiver Operating Characteristic (ROC) curve analysis and calculate the Area Under the Curve (AUC). This will provide a comprehensive view of the model's performance across various thresholds, a crucial step in ensuring that our chosen AI solution meets the rigorous standards for clinical practice and aligns with our commitment to responsible AI deployment in healthcare diagnostics.

The ROC curve for the XGBoost model demonstrates an excellent discriminative performance with an AUC score of 0.86. This score indicates a high degree of accuracy in the model's ability to classify cases correctly, with a substantially better outcome than what would be expected by random chance, as illustrated by the baseline in the curve. The proximity of the curve to the top left corner reflects not only the model's high true positive rate but also a low false positive rate, indicating both high sensitivity and specificity.

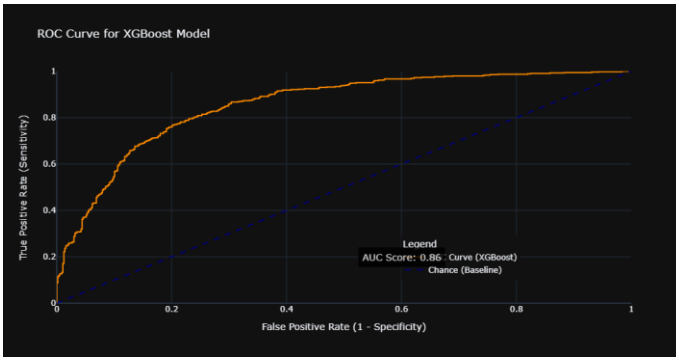

Figure 8: ROC Curve Demonstrating the Discriminative Performance of the XGBoost Model for IDC Diagnosis, with an AUC Score Annotation

This high AUC score is particularly relevant in the clinical setting of breast cancer screening, where the cost of false negatives—missing a diagnosis of cancer—can be extremely high. Similarly minimising false positives, which can lead to unnecessary anxiety and additional medical procedures, is equally important. The AUC score achieved by our model suggests that it strikes an appropriate balance between these two concerns, outperforming the manual inspection rate significantly, and approaching the upper threshold of accuracy presented in systematic reviews (Co et al., 2023).

Conclusions

Throughout our project, we have rigorously evaluated AI models to find a reliable solution for IDC diagnosis within the NHS Breast Cancer Screening Programme. Our chosen model, XGBoost, has demonstrated substantial improvements over manual inspection methods which historically have an

accuracy rate of around 65%. With a sensitivity of 77.21% and specificity of 79.15%, XGBoost approaches the performance levels of AI-assisted methods reported in the systematic review earlier (Co et al., 2023), which found a sensitivity of 83.88% and specificity of 85.49%. Though slightly lower, XGBoost's interpretability and the clarity of its decision-making process offer significant value in a clinical setting, where understanding AI's diagnostic pathways is critical. This balance of high accuracy and transparency positions XGBoost as a favourable alternative to CNNs, which, despite their accuracy, may lack the same level of explanatory power. Our analysis suggests that XGBoost holds the potential to safely augment the precision, speed, and cost-effectiveness of breast cancer diagnostics in healthcare.

Our journey began with the procurement and inspection of a substantial dataset, setting the stage for robust model training. In preprocessing, we employed meticulous techniques to ensure data quality, and through EDA, we gained invaluable insights that informed our model selection, favouring non-linear models due to their ability to handle the dataset's complexities as revealed by UMAP (Figure 4).

The final selection of XGBoost was guided by its slight edge in performance metrics and its suitability for clinical settings, as it offers clarity on diagnostic decision-making. While CNNs also hold potential due to their high accuracy in image classification, their black-box nature poses a challenge for clinical transparency. Therefore, future comparative trials with XGBoost may be beneficial to explore the trade-offs between accuracy and interpretability in depth.

In sum, this project has not only highlighted the potential for AI to revolutionise breast cancer screening but has also underscored the importance of methodical evaluation and the ethical deployment of AI in healthcare. The integration of models like XGBoost into clinical practice holds the promise of enhanced diagnostic processes, provided they are thoroughly vetted for safety, efficacy, and ethical considerations.

References

Co, M., Lau, Y.C.C., Qian, Y.X.Y., Chan, M.C.R., Wong, D.K.K., Lui, K.H., So, N.Y.H., Tso, S.W.S., Lo, Y.C., Lee, W.J., Wong, E., 2023. Artificial Intelligence in Histologic Diagnosis of Ductal Carcinoma In Situ. Molecular and Clinical Pathology Digest, [e-journal]. Available at: https://doi.org/10.1016/j.mcpdig.2023.05.008 [Accessed 20 March 2024].

Dillon, M.F., Quinn, C.M., McDermott, E.W., O'Doherty, A., O'Higgins, N. and Hill, A.D., 2006. Diagnostic accuracy of core biopsy for ductal carcinoma in situ and its implications for surgical practice. J Clin Pathol, 59(7), pp.740-743. Available at: https://doi.org/10.1136/jcp.2005.034330 [Accessed 20 March 2024].

Howard, F.M., Dolezal, J., Kochanny, S., Schulte, J., Chen, H., Heij, L., Huo, D., Nanda, R., Olopade, O.I., Kather, J.N., Cipriani, N., Grossman, R.L. and Pearson, A.T., 2021. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nature Communications, 12(1), Article number: 4423. Available at: https://doi.org/10.1038/s41467-021-24724-0 [Accessed 20 March 2024].

Jung, J., Lee, H., Jung, H., Kim, H., 2023. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. Heliyon, [e-journal] 9(3), e16110. Available through: https://doi.org/10.1016/j.heliyon.2023.e16110 [Accessed 20 March 2024].

Massat, N.J., Dibden, A., Parmar, D., Cuzick, J., Sasieni, P.D. and Duffy, S.W., 2016. Impact of Screening on Breast Cancer Mortality: The UK Program 20 Years On. Cancer Epidemiology, Biomarkers & Prevention, 25(3), pp.455-462. Available at: https://doi.org/10.1158/1055-9965.EPI-15-0803 [Accessed 20 March 2024].

Mooney, P., 2018. Predicting IDC in Breast Cancer Histology Images [online]. Kaggle. Available at: https://www.kaggle.com/code/paultimothymooney/predicting-idc-in-breast-cancer-histology-images [Accessed 20 March 2024].

Taylor-Phillips, S., Seedat, F., Kijauskaite, G., Marshall, J., Halligan, S., Hyde, C. et al., 2022. UK National Screening Committee's approach to reviewing evidence on artificial intelligence in breast cancer screening. The Lancet Digital Health, [e-journal] 4(7), e498-e508. Available through: https://doi.org/10.1016/S2589-7500(22)00088-7 [Accessed 20 March 2024].