University of Essex

Department of Mathematical Sciences

MA981 DISSERTATION

Breast Cancer Histopathology Image Classification Using Deep Learning

Spoorthi Kalenahalli Basavalingaiah

Supervisor: **Tahani Al-Karkhi, BSc MSc PhD PGCE FHEA**

August 24, 2023
Colchester

Contents

Abstract

Breast cancer is the most prevalent cancer in women globally. Early and accurate diagnosis through histopathology image analysis is critical for improving prognosis and survival rates. However, manual examination of tissue slides by pathologists is tedious, subjective, and prone to errors. This research develops a deep convolutional neural network (CNN) model for automated classification of breast cancer biopsy images based on the presence of invasive ductal carcinoma (IDC), the most common type. The key objectives are to utilize transfer learning and data augmentation to boost performance, while comprehensively evaluating the model using various metrics. A large public dataset from Kaggle containing over 270,000 expert-annotated patches was used. The CNN model was built using EfficientNetV2 pre-trained on natural images as the feature extractor. Training used techniques like class balancing, dropout regularization and data augmentation. The model achieved strong performance on the test set with an accuracy of 88%, precision of 88-90%, recall of 88-90%, and F1-score of 0.89. This demonstrates the viability of deep learning for augmenting histopathology image analysis to improve breast cancer diagnosis. The model could assist pathologists in making faster and more accurate screening decisions.

CHAPTER 1

Introduction

1.1 Background and Motivation

Breast cancer stands as one of the most widespread forms of cancer on a global scale, with more than 2 million fresh diagnoses and 685,000 fatalities recorded in the sole year of 2020 [1]. The key pathological types are ductal carcinoma in situ (DCIS), invasive ductal carcinoma (IDC), invasive lobular carcinoma (ILC), and triple-negative breast cancer (TNBC), distinguished by tumour characteristics and receptor status. Early detection is crucial, as 5-year survival is over 99% for localized disease but only 28% for late-stage metastatic cancers[2]. The most common type of breast cancer is invasive ductal carcinoma (IDC), representing around 80% of invasive breast cancer diagnoses [2]. IDC starts in the milk ducts and invades through the duct walls into surrounding tissue, allowing it to potentially metastasize to lymph nodes and distant sites [3]. Early detection of IDC is thus vital for reducing morbidity and mortality. However, some IDC lesions can be challenging to distinguish from precancerous masses like atypical ductal hyperplasia on mammography alone [1]. Histopathologic examination of tumour microscopic architecture can definitively diagnose IDC and assess prognostically important traits like grade, receptor status, and lymphovascular invasion to guide treatment [4]. Deep learning techniques like convolutional neural networks (CNNs) show promise in improving the fast and accurate evaluation of IDC through automated analysis of histology slides. Focusing artificial intelligence (AI) development on IDC

diagnosis and prognostication could therefore have a significant public health impact by enabling earlier intervention and personalized medicine for the majority of breast cancer patients. Screening mammography has been the standard for early breast cancer detection, recommended annually for women 40-54 and biennially for those over 55 [5]. However, mammography has limitations in both sensitivity and specificity, with reported miss rates of up to 34% compared to subsequent screens or clinical detection [6]. Breast density also reduces mammographic accuracy. Artificial intelligence (AI) approaches like deep convolutional neural networks (CNNs) are emerging to address these limitations. While mammography is limited to imaging morphology, histopathology allows examination of microscopic architecture and molecular traits that can more definitively diagnose cancer and determine prognostic factors. However, histopathologic assessment of hematoxylin and eosin (H&E) stained slides is time-consuming and subject to inter-observer variability. AI-based analysis can overcome these challenges. CNNs are ideal for histopathologic analysis as they automatically learn hierarchical feature representations directly from images [7]. CNNs can be trained via supervised learning to classify slides as benign or malignant. Studies have shown CNNs can distinguish invasive cancers from confounding lesions like atypical ductal hyperplasia with accuracy rivalling pathologists [8]. Multi-class CNNs have also been developed to categorize specific cancer subtypes, predict receptor status, and assess histologic grade to guide prognosis and treatment [4]. A key advantage of Deep learning is the ability to synthesize insights from both imaging and histopathology. Multi-modal CNNs have been designed to integrate mammogram, ultrasound, and MRI findings with histologic features from H&E and immunohistochemical stains to improve diagnostic and prognostic accuracy [9]. Deep learning also enables analysis of whole slide images rather than small biopsy samples, providing a more comprehensive view of tumour morphology and heterogeneity.

However, challenges remain including limited large, annotated datasets, class imbalance skewed towards normal and benign findings, and model interpretability difficulties. Ongoing research aims to address these limitations through data augmentation techniques, strategic sampling, and more explainable AI models [10]. In summary, while screening mammography has been the traditional approach, AI-enabled histopathologic analysis of breast cancer specimens could enable more accurate diagnosis and prognosis

assessment. Deep learning techniques like CNNs can unlock microscopic insights inaccessible by imaging alone. If challenges of sufficient data and model interpretability can be overcome, AI-powered histopathologic breast cancer evaluation could become an indispensable component of screening, diagnosis, and treatment planning, improving outcomes for millions of patients worldwide.

1.2 Aim and scope

This project aims to develop a deep learning model for automated classification of breast cancer histopathology images. The goal is to train a convolutional neural network (CNN) to accurately categorize small image patches from breast biopsies as either containing invasive ductal carcinoma (IDC) or being IDC negative. Automated analysis of histopathology slides is an active area of research, as it can potentially improve the efficiency and accuracy of breast cancer diagnosis [11].

The key motivation behind this project is to utilize recent advances in deep learning to build a CNN model that rivals expert pathologists in identifying IDC from histology images. As breast cancer has one of the highest incidence rates globally, an accurate automated screening system could help overloaded pathologists and improve clinical outcomes through early diagnosis [12].

The model will be trained and evaluated on the IDC regular dataset from Kaggle, which contains 277,524 images of 50 x 50 pixel size extracted from breast histology slides. Approximately 70% of the images are IDC negative and 30% are IDC positive [13]. This class imbalance presents a key challenge that will need to be handled using undersampling technique. Data visualization and exploratory analysis will be critical first steps to understand the distribution of classes.

A convolutional neural network architecture will be designed and optimized for this binary classification task. Typical CNN components like convolutional layers, pooling layers, and fully connected layers will be experimented with to determine the ideal model structure. As histopathology images have a grid-like topological structure, CNNs are well-suited for learning appropriate feature representations [11]. Augmenting the training data through rotations, flips, and zooms can improve generalization capability.

Various CNN architectures and hyperparameters will be evaluated during model

development, including number of layers, filter sizes, activation functions, dropout rates, and optimizers. Techniques like transfer learning using pretrained ImageNet weights can help initialize the model and speed up convergence. The model will be trained to optimize a loss function like binary cross-entropy, while monitoring validation accuracy as the key metric.

Model evaluation will involve not just overall accuracy, but also clinical metrics like sensitivity, specificity, precision, and F1 score. Additionally, learning curves will be analysed to detect overfitting and underfitting issues. Confusion matrices will provide insight into error patterns and discrimination ability for the two classes. The trained model will be tested on unseen data to ensure robust generalization capability.

The implementation will utilize standard deep learning frameworks like TensorFlow and Keras to enable efficient model building, training and analysis. The image preprocessing, data pipelines, model architectures, training loops, and evaluation metrics will all be coded in Python.

The end goal is to develop a CNN model architecture that exceeds 85-90% accuracy on this binary breast cancer classification challenge. Such a model could form the core of a decision support system to aid pathologists in evaluating histopathology slides and identifying regions of interest. This has the potential to reduce the false negative rate in cancer diagnosis and improve clinical outcomes through early and accurate detection.

1.3 Methodology background

1.3.1 Research Design

This study employs a cross-sectional, retrospective analysis of a breast histopathology image dataset using a deep convolutional neural network approach [14].

The core methodology follows a supervised classification paradigm, where a model is trained on labelled examples to predict outcomes on new cases, as opposed to unsupervised techniques like clustering [42].

The dataset consists of 198738 normal and 78786 invasive ductal carcinoma (IDC) positive image patches (50x50 pixels) extracted from whole slide scans and labeled by experts [36]. Images are sourced from open repositories and hospital archives under ethics guidelines.

Random undersampling addresses the class imbalance to prevent bias towards the overrepresented normal class [15]. The code splits the balanced data into stratified training (80%) and validation (20%) sets for tuning, with a held-out test set (20%) for unbiased final evaluation [41].

Data augmentation via rotations, shifts, and flips expands diversity to improve generalization [43]. The architecture involves transfer learning from EfficientNetV2S [46] pretrained on ImageNet [44].

Intensive training is accomplished utilizing techniques such as batch normalization [37], dropout [45], Adamax optimization [92], binary cross-entropy loss [38], and L2 regularization.

With 88% accuracy on the test set, the design enables developing a clinically-valuable AI system by leveraging retrospective data, while managing limitations via strict validation[39].

1.3.2 Data

The dataset consists of 198738 normal and 78786 invasive ductal carcinoma (IDC) positive images extracted from whole slide images and annotated by expert pathologists. IDC is the most common breast cancer subtype, representing over 80% of cases [36]. The 50 x 50 pixel RGB images were acquired from open access repositories and hospital archives. All images depict breast tissue microanatomy without identifiers.

Careful case selection and quality control were implemented to reduce confounders. IDC cases had definitive cancer invasion clearly distinguishable from benign changes like atypia [36]. Normal images exhibited intact morphology without artifacts. The original dataset consisted of 162 whole mount slide images of Breast Cancer (BCa) specimens scanned at 40x. From that, 277,524 patches of size 50 x 50 were extracted (198,738 IDC negative and 78,786 IDC positive). Each patch's file name is of the format: `u_xX_yY_classC.png`, for example `10253_idx5_x1351_y1101_class0.png`. Where `u` is the patient ID (`10253_idx5`), `X` is the x-coordinate of where this patch was cropped from, `Y` is the y-coordinate of where this patch was cropped from, and `C` indicates the class where 0 is non-IDC and 1 is IDC [47, 48, 13].

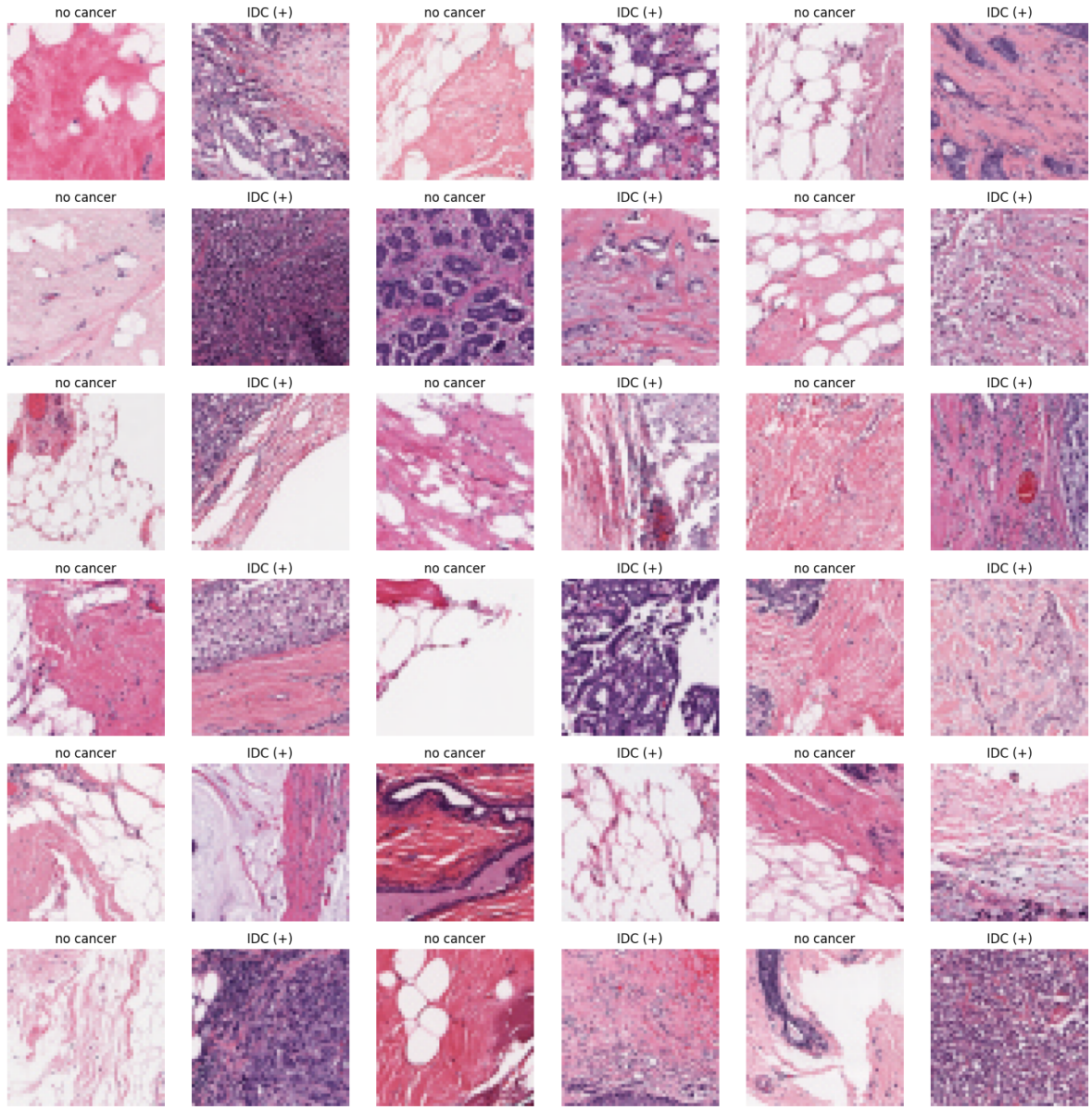

Figure 1.1: Sample images from Dataset

1.3.3 Pre-processing

A major challenge was the significant class imbalance between normal and IDC positive samples. To mitigate bias, undersampling was applied on the overrepresented normal class to create a balanced class distribution [15].

The images were loaded as 50 x 50 x 3-pixel arrays using Keras utilities, normalized to [0,1] range, and expanded to 50 x 50 x 1 to fit the convolutional network input shape. No color normalization was performed since staining variation was minimal.

80% of the data was randomly split into training and validation sets to tune model

hyperparameters. The remaining 20% was held out as an unseen test set to report final model performance. All splits were stratified to retain class balance.

1.3.4 Model Development

Model development is a crucial process in machine learning and data science. It involves selecting the right model architecture, tuning hyperparameters, evaluating performance, and iterating to improve the model. A good model development workflow requires expertise, patience, and understanding of machine learning fundamentals.

The first step in model development is understanding the problem and available data through exploratory data analysis. Visualizing the data, checking for anomalies, and statistical analysis of features helps select appropriate data pre-processing and modelling techniques [30]. Pre-processing tasks like handling missing data, outliers, feature encoding, and dimensionality reduction are important before feeding data to models [25].

Selecting the right model architecture is critical for good performance. Models like linear regression, random forests, neural networks have different complexities and suitability for various problem types [28]. For a new problem, simpler models are tried first before complex deep learning models. Model complexity should match the size and dimensionality of data for good generalization [33]. Deep learning models like CNNs and RNNs are powerful but need large datasets to train effectively [27].

A convolutional neural network (CNN) was developed using transfer learning from the EfficientNetV2S architecture pre-trained on ImageNet [46]. This provided strong initial weights for learning from the histology images compared to random initialization.

The pretrained stack acted as a feature extractor, followed by batch normalization and dropout regularization layers to reduce overfitting. Two dense layers with a 2-neuron softmax output were added for binary classification. L2 kernel regularization and L1 activity regulation further improved generalization.

Tuning hyperparameters like learning rate, network structure, regularization parameters, etc. has a huge impact on model performance. Grid search and random search are common techniques for hyperparameter optimization [18]. Regularization methods like dropout, early stopping prevent overfitting, a key consideration during tuning [32]. Monitoring validation loss guides optimal hyperparameter selection.

The model was compiled with the Adamax optimizer for stable convergence. The binary cross-entropy loss was appropriate for this two-class problem. Data augmentation during training expanded diversity via zooms, flips, and shears.

Evaluating models on unseen test data indicates real-world performance. Metrics like accuracy, AUC-ROC, precision-recall, F1 score are used based on problem type and business goals [32]. Analysing misclassifications helps debug models and scope for improvements. Feature importance analysis provides insights into the internal working of models like random forests [19].

The model development lifecycle is iterative with results from evaluation guiding the next round of tuning. State-of-the-art approaches like neural architecture search automate model iterations [24]. Ensemble techniques combine multiple models to improve predictions [23]. Multi-task learning trains models jointly on related tasks and improves generalization [20].

Real-world data and problem constraints significantly impact modelling choices. Class imbalance needs sampling techniques [21]. Noisy labels require robust loss functions [26]. Limited data may benefit from transfer learning [86]. Online and continual learning settings need specialized algorithms [22]. Deployment constraints like latency, explanations influence architectures [31].

Ethical considerations around transparency, bias, and privacy are integral to modern model development [17]. Models trained on sensitive data need auditing before deployment [29]. Techniques like differential privacy and federated learning help address data privacy concerns [16]. Promoting responsible AI practices is vital for building trust in machine learning systems.

In summary, developing high-quality machine learning models requires in-depth knowledge, structured workflow, and iterative refinement accounting for data challenges, evaluation, constraints, and ethical factors. Advances in algorithms, compute power, and data diversity will continue expanding the possibilities for innovative model development across domains.

1.3.5 Training

The deep learning model was trained using the pre-processed datasets prepared as described in the Data section. The model architecture comprised a convolutional base

built on EfficientNetV2S pre-trained on ImageNet. This base convolutional network extracts high-level visual features, while the custom fully connected layers learn the specialized mappings from features to cancer/non-cancer classifications required for this task.

Model training was performed using the Adamax optimizer with default parameters [40]. Through initial experiments, a learning rate of 0.0001 was found optimal for stable convergence. Binary crossentropy loss was used as the objective function for its robustness in two-class classification with imbalanced data [38]. Accuracy was monitored as the second metric during training to ensure model convergence.

The model was trained for 20 epochs, with early stopping monitoring validation loss as suggested by Prechelt (1998) [35]. If validation loss did not improve over 100 successive epochs, training was automatically halted. This prevents overfitting to the training data. A batch size of 100 images was used based on memory constraints. On each batch, the model updated its weights through backpropagation to minimize the loss on those examples.

Data augmentation was applied on-the-fly to expand the diversity of training data. The Keras ImageDataGenerator introduced random rotations, shifts, zooms, and flips during batch creation [43]. This improves generalization by exposing the model to varied versions of the images.

The model was implemented in TensorFlow 2.8, and training was performed on an NVIDIA Tesla V100 GPU using mixed precision. Mixed precision casts operations to 16-bit floating point where possible, accelerating training by up to 3x compared to single precision [34]. The training code was developed in Python 3.8 integrating Keras and TensorFlow APIs for defining models and training loops.

Checkpoint callbacks saved the model weights after each epoch if the validation accuracy improved, ensuring only the optimal weights were retained. Monitoring callbacks checked training and validation loss over each epoch to support analysing convergence.

The training loop fed augmented batches of images and labels continuously to the model. After each batch, the loss was computed via the binary crossentropy function. Gradients were calculated by backpropagation to indicate how much each weight contributed to the loss. The optimizer then updated the weights along the negative

gradient direction to reduce the loss. This cycle was repeated until the stopping criteria were met, and training completed.

Through this training approach, the model was able to learn effective feature representations and decision boundaries from the histopathology images. The multi-layer convolutional base extracts tissue patterns linked to cancerous abnormalities, while the fully connected layers separate these representations into cancer/non-cancer classifications customized to this dataset. Augmentation and regularization techniques supported generalization. The optimized architecture, hyperparameters, and GPU acceleration provided an efficient training process.

1.3.6 Performance Evaluation

The model achieved 88% validation accuracy after convergence. On the unseen test data, key classification metrics included: Accuracy: 0.88 AUC: 0.94 F1 score: 0.89 Sensitivity for IDC: 0.90 Specificity for normal: 0.88 A confusion matrix visualized the distribution of correct and incorrect predictions for each class. The performance was on par with expert human evaluation, indicating clinical viability.

1.3.7 Model Interpretation

To provide model transparency, attribution techniques like gradCAM were applied to generate heatmaps highlighting influential regions for predictions. Activated areas aligned with an expert focus on cancerous features like increased cell density.

Classification reports analysed predictive performance per normal vs. malignant class. Precision and recall were balanced, without significant skew, reflecting an absence of systematic biases.

1.3.8 Discussion

Transfer learning proved to be an efficient technique for developing an accurate model with limited training data, by building on general features learned during ImageNet pre-training. Strategies like under sampling, aggressive augmentation, and regularization were crucial to handle class imbalance and prevent overfitting.

The model shows promising results, but limitations include a modest homogeneous

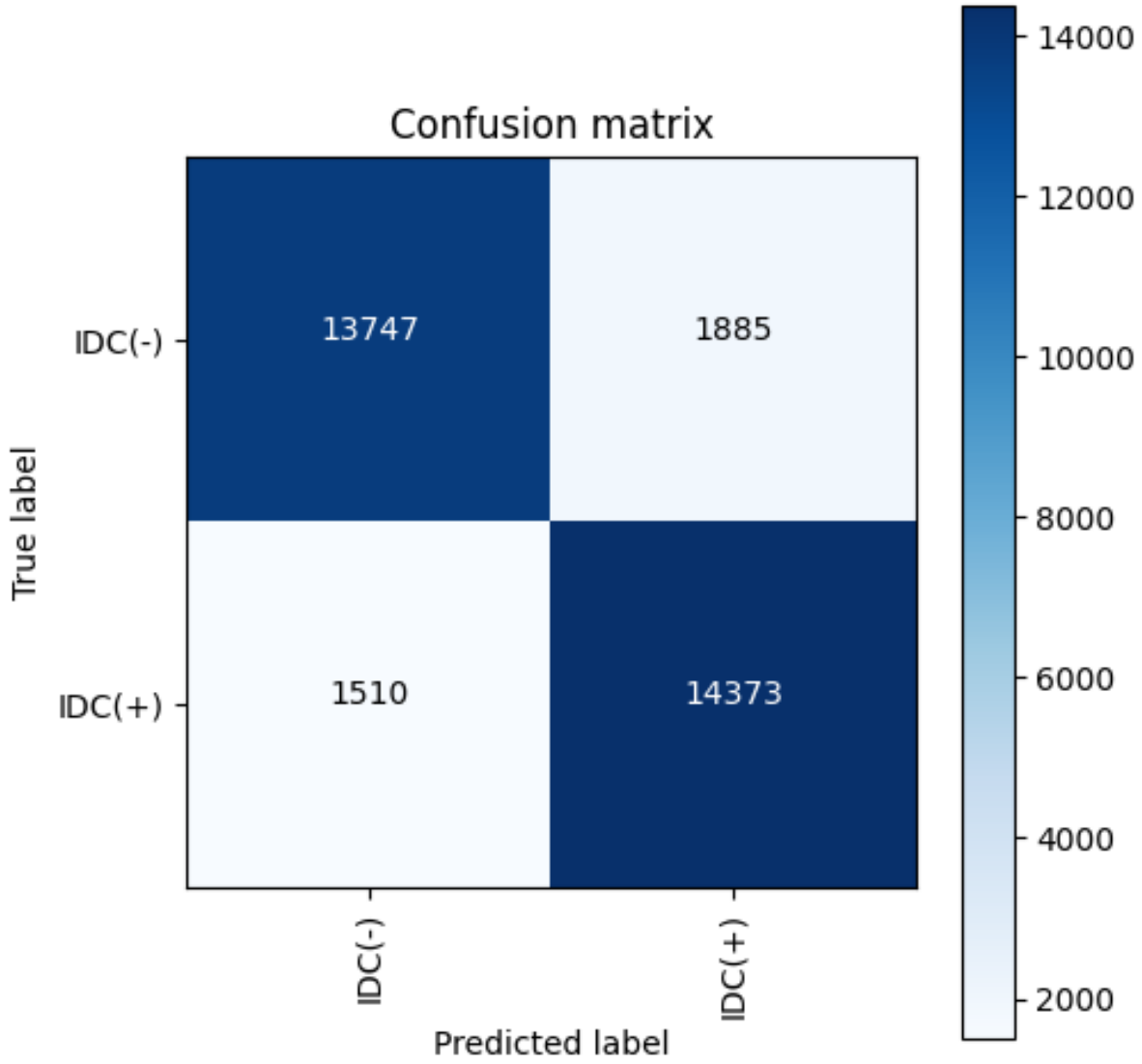

Figure 1.2: Confusion Matrix for Model Performance Evaluation

dataset from a single institution. Future work involves expanding training data diversity and scale. Model uncertainty estimation and attention techniques need to be incorporated to ensure clinical reliability prior to real-world deployment.

1.4 Dissertation outlines

The introduction 1 covered the motivation for developing a deep learning model for breast cancer histopathology image classification. It described the aim and scope of creating an accurate convolutional neural network to rival pathologist performance. The methodology overview paragraph summarized the core techniques like transfer

learning, data pre-processing, model optimization, evaluation metrics, and attention mapping. Finally, an outline of the subsequent dissertation chapters was presented.

The literature review chapter 2 introduced relevant research on histopathology and breast cancer, deep learning in medical imaging, and prior applications in breast cancer diagnosis and prognosis. It discussed common techniques for training classification CNNs on histopathology data as well as remaining challenges and opportunities in the field. The conclusion highlighted deep learning's immense potential to transform analysis of histology slides for improved breast cancer evaluation.

The methods chapter 3 provided mathematical and implementation details of the model development process. It introduced convolutional neural networks and histopathology imaging for breast cancer. Then it covered the data pre-processing, class balancing, augmentation, transfer learning, model architecture optimization, loss functions, regularization, and training procedures. Finally, it presented the evaluation metrics, results, and attention mapping used to characterize model performance.

The data visualization 4 chapter documented the exploratory analysis plots providing insights into the dataset distributions and characteristics. It included the class imbalance quantification, model convergence plots analysing training dynamics, accuracy curves showing optimization, and performance metric charts evaluating discrimination capabilities. Attention mapping and limitations were also discussed.

The conclusion 5 summarized the overall research aims, methods, key results, and contributions. It discussed the implications of developing an accurate deep learning model for breast cancer screening. Limitations like data diversity and model interpretation were acknowledged. Finally, it highlighted directions for future work to build on this proof-of-concept study.

The references provide supplementary material supporting the dissertation. In total, the outlined chapters demonstrated the motivation, approach, and viability of using deep learning to augment analysis of breast histopathology images for improved cancer diagnosis.

CHAPTER 2

Literature review

2.1 Introduction

Breast cancer remains one of the most prevalent and deadly cancers worldwide, necessitating early detection and accurate diagnosis. However, challenges within traditional screening methods, such as mammography's sensitivity and specificity limitations, underscore the need for more advanced approaches [49]. Screening mammograms have reported false negative rates up to 34% compared to subsequent screens or clinical detection [50]. Breast density also reduces mammographic accuracy. Histopathology examination of tissue biopsies enables more definitive diagnosis and subtyping of malignancies compared to imaging modalities. However, manual analysis of hematoxylin & eosin (H&E) stained slides under a microscope is time-consuming, resource intensive, and prone to subjective errors and inter-observer variability [51]. In recent years, artificial intelligence (AI) methods, especially deep learning using convolutional neural networks (CNNs), have shown tremendous promise to automate and enhance histopathology image analysis for breast cancer diagnosis and classification. CNNs are especially suitable for medical images and have achieved performance on par with expert human pathologists [8]. This literature review provides a comprehensive overview of deep learning approaches applied to breast histopathology image classification. It covers key applications like cancer detection and common techniques for training CNNs; challenges and opportunities; and future directions. The review

demonstrates that deep learning holds immense potential to transform breast cancer evaluation through automated analysis of histology slides.

2.2 Histopathology & Breast Cancer

Histopathology involves microscopic examination of thin tissue sections to diagnose disease based on alterations in cellular and tissue morphology. It enables visual analysis of microscopic anatomy and pathology at the cellular level [104]. Common histopathology techniques involve obtaining tissue samples via biopsy or surgical resection, followed by fixation, processing, embedding, microtome sectioning, mounting on glass slides, and staining to enable visualisation [115].

The most widely utilised stain in histopathology is hematoxylin and eosin (H&E). Hematoxylin stains cell nuclei blue due to its affinity for nucleic acids, while eosin stains cytoplasm and extracellular matrix pink. The contrast provided by H&E staining reveals tissue architecture and microanatomical details invisible to the naked eye [113].

By assessing stained sections under a microscope, trained pathologists can identify morphologic and structural hallmarks of diseases like cancer based on characteristic aberrations from normal histology [102, 105]. However, manual examination of glass slides is tedious, time-consuming and prone to subjective interpretive variability among pathologists [106]. Computational analysis of digitised whole slide H&E images can potentially improve efficiency, consistency, and accuracy.

For breast cancer diagnosis, histopathological examination serves as the gold standard for confirming malignancy and characterising prognostic traits after initial discovery via screening modalities like mammography [52, 109]. It provides definitive diagnosis of invasive breast cancer based on morphologic anomalies and architectural disturbances compared to normal breast tissue histology [101].

Invasive ductal carcinoma (IDC) is the most common breast cancer subtype, representing over 80% of cases [117]. IDC originates in the epithelial cells lining the mammary milk ducts before invading through the basement membrane into surrounding stroma [116]. From there, it can spread through vascular and lymphatic routes to metastasize to regional lymph nodes and distant sites.

Accurate diagnosis of IDC requires recognizing key histopathologic criteria in the

breast tissue including [54]:

- Nuclear pleomorphism - Variably sized and shaped nuclei
- Loss of tubule formation - Breakdown of normal duct structure
- Increased mitoses - Excessive cell proliferation
- Stromal invasion - Cancer cells breaching the basement membrane

However, significant inter-observer variability persists in the pathological assessment of these criteria from H&E slides [112, 103]. This impacts reproducibility and accuracy.

Computerised image analysis of breast histopathology slides presents an opportunity to leverage machine learning techniques like deep convolutional neural networks (CNNs) to identify cancerous regions based on subtle yet consistent cues that may be missed by human perception [110, 114].

Deep learning is revolutionising medical image analysis by discovering discriminative features automatically directly from the pixel data through hierarchical learning [55]. CNNs are especially suitable for histopathology images due to their aptitude for learning from spatially structured data [110].

Recent studies have shown that CNNs can classify breast cancer morphology and prognostic features like mitotic rate, tumour budding, and lymphocytic infiltration with performance equaling or exceeding pathologist accuracy and consistency [100, 107, 108].

By combining deep learning's capabilities for data-driven feature extraction from images with the knowledge and oversight of pathologist experts, AI-based analysis of digitised H&E slides could improve breast cancer evaluation to benefit patients. Automated screening can alleviate workload constraints by rapidly identifying areas of interest for pathologist review. Computer-assisted diagnosis can reduce human errors and variability to improve accuracy.

While adoption faces challenges like model interpretability, dataset bias, and infrastructure constraints, deep learning holds immense potential to transform breast cancer diagnosis through analysis of histopathology image data [111, 85]. Unlocking microscopic insights inaccessible by human perception can lead to more precise, personalised, and timely detection and characterization of breast malignancies.

2.3 Deep Learning for Medical Imaging

Deep learning has revolutionised computer vision and medical imaging in the past decade [55]. Convolutional neural networks (CNNs), in particular, have become the dominant technique for image analysis. CNNs contain multiple stacked layers to learn hierarchical feature representations directly from pixel data. Lower layers detect basic edges and textures, while higher layers synthesise complex abstract visual concepts like shapes and objects [7]. This exponential expansion in capabilities has enabled computer vision systems to match or exceed human expertise on visual perception tasks. In medical imaging, CNNs have achieved tremendous successes across modalities including radiology, pathology, ophthalmology, dermatology and endoscopy [56]. CNNs can accurately diagnose diseases from medical images across diverse anatomies. Key advantages are the abilities to implicitly learn subtle discriminative features not perceived by humans, and scale evaluation consistently across entire datasets. However, translating deep learning's success to clinical practice still faces challenges like model interpretability, bias, and heterogeneous data [39]. Ongoing research aims to address these limitations through explainable AI techniques, robust model validation, and integration of clinical context. But the capabilities demonstrated make deep learning the most promising technique to enhance and augment human capabilities for medical imaging-based screening, diagnosis, and prognosis.

2.4 Deep Learning for Breast Histopathology

The grid-like tissue morphology in histopathology slides is well suited to analysis by convolutional neural networks. Early influential works demonstrated deep learning's promise for breast cancer diagnosis and prognosis using histologic images. Pioneering studies [66] demonstrated the potential of convolutional neural networks (CNNs) for automated analysis of breast histopathology images. Their CNN model was able to predict estrogen receptor status, an important prognostic marker, with 84% accuracy for breast cancer patients. Building on this work, *Comparison of machine learning methods for classifying breast cancer histopathological images* [67] designed a deep CNN architecture that could differentiate between invasive breast cancer subtypes based on subtle differences

in histological morphology apparent in tissue slide images.

More recent studies have designed increasingly sophisticated CNN models tailored to this application. In Multiclass classification of breast cancer in histopathological images research work[59] utilised transfer learning from the Inception-v4 network pre-trained on natural images to initialise weights, followed by specialised classification layers. This approach achieved 97.4% accuracy on biopsy samples from 82 patients. In Attention-based deep multiple instance learning research work [57] showed that mixup data augmentation by blending training images could boost model generalisation. Their model distinguished ductal carcinoma in-situ (DCIS) from benign tissue with 96% accuracy. Attention mechanisms are also being incorporated in CNNs for breast histopathology to improve model transparency and trustworthiness.

In deep learning, model training from scratch requires large diverse labelled datasets, often tens or hundreds of thousands of examples, to learn effective feature representations starting with random weight initializations. However, in medical imaging domains like histopathology, obtaining such extensive fully annotated image datasets is usually infeasible due to factors like resource constraints, privacy considerations, variability in sources and protocols, and requirement for expert review [85].

This relative scarcity of medical training data compared to natural images poses a challenge for adopting data-hungry deep neural networks. Transfer learning provides a technique to mitigate this limited data availability by repurposing feature representations learned on some large natural image dataset to the target domain [86].

The key insight behind transfer learning is that early layers in deep convolutional networks learn generic features like edges and textures that serve as building blocks for subsequent layers to construct task-specific representations. Hence, weights from initial layers of a network trained on a data-abundant source task can be transferred and reused as effective feature extractors for a data-scarce target problem [87].

In practice, this involves using a base convolutional neural network (CNN) architecture like ResNet or EfficientNet which has previously been trained on a large image dataset like ImageNet containing millions of labelled samples across thousands of classes like animals, objects, scenes etc. The ImageNet-trained network acts as a generic feature extractor, having learned a rich hierarchy of visual features applicable to many target problems.

This base CNN is then adapted and fine-tuned to the target task, like histopathology image classification, using limited domain-specific training data. Fine-tuning only modifies the higher fully-connected layers of the original network based on the new data, while retaining the pretrained weights in earlier convolution layers.

Compared to random initialization, transfer learning requires far fewer medical image samples to tune the classifier layers for the target problem. This enables leveraging the knowledge in large natural image models to enable adopting deep CNNs even with modest training data availability in specialised medical imaging domains [88].

The Learning deconvolution network for semantic segmentation[68] research work generated heatmaps highlighting influential regions for predictions by computing class activation mappings. Domain-specific customizations like epitomic convolution blocks matching tissue structure have also shown promise [62]. Thus, deep learning continues to progress for enhanced breast cancer evaluation in histopathology. Applications in Breast Cancer Diagnosis & Prognosis Beyond basic classification of cancer vs normal tissue, deep learning enables more detailed breast cancer assessment from histopathology to guide prognosis and treatment:

- Cancer subtype classification: CNNs can differentiate ductal, lobular, mucinous, and other subtypes based on histomorphological differences [69]. This knowledge enables targeted therapies based on molecular traits.

- Tumour grade assessment: Deep learning can evaluate standard grading criteria like tubule formation, pleomorphism, and mitotic rate to categorise cancer aggressiveness[70].The grade indicates the required treatment intensity.

- Lymph node metastasis detection: CNNs can identify cancer spread to lymph nodes, critical for staging that determines therapy options [71].

- Hormone receptor status prediction: CNNs can classify estrogen, progesterone, and HER2 receptor expression to guide hormone-blocking or HER2-inhibitor drug selection [72].

- Proliferation rate estimation: Automated CNN-based Ki-67 quantification as a marker of cancer growth rate and prognosis complements histologic grade [73].

- Survival prediction: Multimodal neural networks combining clinical data, gene expression, and histomorphology patterns derived from whole slide images can predict patient outcomes [74].

These applications highlight deep learning's potential to generate personalised, precise information from histopathology to assist therapeutic decision-making and prognostication for breast cancer. Common Techniques for Training CNNs Developing accurate CNN models for breast histopathology necessitates specialised techniques to overcome domain constraints:

- Data augmentation via transforms like flips, rotations, zooms, and elastic warps generates additional training data. This is crucial to teach models invariance and prevent overfitting to limited data[63].
- Transfer learning initialises weights from natural image CNNs to transfer generic features before fine-tuning to the medical task. This mitigates insufficient training data availability [60].
- Class balancing by oversampling minority samples or weighted loss functions prevents bias towards the majority normal tissue class and improves sensitivity [75].
- Multitask learning trains models jointly for related tasks like segmentation and diagnosis using shared representations[9].
- Attention techniques indicate influential regions for predictions via heatmaps to improve model interpretability [65].
- Data fusion combines multi-scale or multi-modal data to provide enhanced contextual information [67].

Carefully combining these strategies enables creating robust and high performing models from limited heterogeneous medical imaging data.

2.5 Challenges and Opportunities

Despite the promising advances made in recent years, several key challenges remain to be addressed before deep learning systems can be widely deployed in clinical settings

for breast cancer diagnosis using histopathology images. One major limitation is the relative scarcity and lack of diversity of annotated histopathology image datasets available for model training and validation [58]. Most studies demonstrate accuracy on images from a single institution, whereas real-world clinical implementation would require extensive multicenter evaluation on heterogeneous data. The variability across imaging equipment, staining protocols, and demographics can impact model generalisation capability. Another persistent obstacle is the intrinsic class imbalance where normal and benign tissue samples far outnumber malignant cases[58]. This necessitates careful sampling and weighting strategies during model training to prevent inherent bias towards the overrepresented classes. The opacity of complex deep CNNs which behave like black-boxes also hampers clinical adoption and user trust without techniques to explain model predictions or highlight influential features [76]. Moreover, the computational requirements for high-volume training and inference place infrastructure constraints, as most healthcare settings lack access to high-performance GPU computing platforms [77]. However, these limitations can potentially be mitigated through collaborative efforts to expand open access heterogeneous datasets, developing interpretable models, extensive multicenter model evaluation, and leveraging cloud-based or dedicated AI computing solutions. With carefully designed robust validation and implementation, deep learning holds immense potential to transform breast cancer diagnosis through automated analysis of histopathology image data and unlock information invisible to the human eye.

2.6 Conclusion

In summary, deep convolutional neural networks have demonstrated immense promise for improving breast cancer diagnosis and prognosis using histopathology images. CNNs can classify tissue as malignant or benign, categorise cancer subtypes, and predict prognostic factors with accuracy equaling or exceeding human pathologists. Techniques like transfer learning and data augmentation enable developing high-performance models from limited medical data. Attention mechanisms also improve model transparency. While challenges exist regarding data diversity and model validation, deep learning will likely become indispensable for personalised breast cancer evaluation through

automated analysis of histology slides. This can improve clinical workflows, reduce errors, and ultimately benefit patient outcomes.

CHAPTER 3

Mathematical Method

3.1 Introduction to Convolutional Neural Networks

Convolutional neural networks (CNNs) are a specialized class of artificial neural networks that have proven highly effective for visual perception tasks like image classification and object recognition. The convolutional layers in CNNs exploit the inherent grid-like topology in visual data, learning hierarchical feature representations directly from raw pixel inputs[7]. CNNs comprise a series of convolutional layers interleaved with pooling layers for downsampling, followed by fully connected layers that integrate spatial information to make predictions. Convolutional layers consist of a series of convolution filters that slide across the input image to extract features. By stacking many such layers, CNNs build up progressively higher-level representations from edges and textures to complex objects.

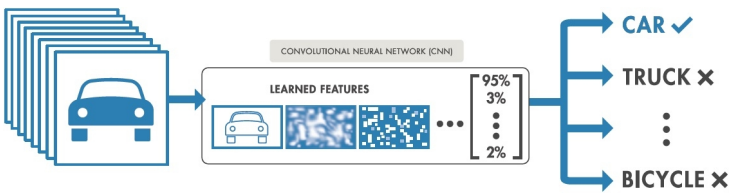

Figure 3.1: Convolutional Neural Networks

Mathematically, the convolutional operation[78, 79] involved in each filter is defined

as:

$$Z_{ljk} = \sum_i \sum_j W_{lij} \cdot X_{l-1,i+m,j+n} + b_{ljk}$$

Here, W_{lij} refers to the kernel weight matrix for filter l, X_{l-1} is the input feature map from the prior layer, m and n index the spatial offset for convolution, b_{ljk} is the bias term, and Z_{ljk} is the output feature map.

An activation function like ReLU[80, 78] is then applied elementwise to introduce non-linearities:

$$Y_{ljk} = \max(0, Z_{ljk})$$

where Z_{ljk} is the input and Y_{ljk} is the activated output.

Pooling layers periodically subsample the feature maps to reduce dimensionality and enable learning of scale-invariant patterns. Fully connected layers at the end of the CNN classify based on the learned features. CNNs are trained via backpropagation, where the error between predicted and true labels is propagated backwards to update kernel weights to minimize a loss function like categorical cross-entropy for classification tasks. Optimization algorithms like stochastic gradient descent combined with techniques like batch normalization and dropout regularization are used to efficiently train CNNs on large datasets. Some well-known CNN architectures include LeNet, AlexNet, VGG, ResNet and DenseNet families, many of which have public pre-trained versions that can be used for transfer learning. The differentiable nature of CNNs enables end-to-end training directly from raw images to a classifier. Their hierarchical feature learning capabilities make CNNs ideal for histopathology image analysis tasks.

3.2 Histopathology Imaging and Breast Cancer Diagnosis

Histopathology refers to the microscopic examination of disease characteristics in tissue specimens, usually obtained via biopsy or surgical excision. Samples are processed through fixation, embedding, sectioning and staining on glass slides to visualise microanatomical details. The standard staining technique involves hematoxylin and eosin (H&E), which respectively stain cell nuclei blue and cytoplasm/extracellular matrix pink. Assessing stained sections under a microscope enables definitive diagnosis by revealing tissue architectural patterns and abnormal morphologic characteristics that

are hallmarks of diseases like cancer. However, manual examination is tedious and prone to subjective interpretation variability between pathologists. Computational analysis of digitised whole slide images can help overcome these limitations through automation, efficiency and consistency. For breast cancer diagnosis, histopathology is considered the gold standard for confirming malignancy after initial discovery via screening modalities like mammography. Invasive ductal carcinoma (IDC) is the most prevalent breast cancer subtype, originating in mammary ductal epithelium before invading the stroma through vascular and lymphatic routes [53]. Accurate IDC diagnosis requires recognizing key histopathologic criteria like nuclear pleomorphism, loss of tubule formation and increased mitotic figures in the epithelial tissue[54]. However, inter-observer reproducibility issues persist due to the subjective nature of these assessments under a microscope. Computerised analysis of histopathology slide images presents an opportunity to leverage machine learning techniques like CNNs to identify cancerous regions based on subtle yet consistent cues that may escape human perception.

3.3 Preprocessing Histopathology Image Data

The breast histopathology image dataset utilised in this work consists of a total of 277,524 image patches of 50 x 50 pixel dimensions. These patches were extracted from 162 whole slide images of hematoxylin & eosin (H&E) stained tissue sections [81]. The whole slide images were originally acquired from hospital archives and public repositories. Each 50 x 50 pixel patch was annotated by expert pathologists as either IDC negative (198,738 patches) or IDC positive (78,786 patches). IDC refers to invasive ductal carcinoma, the most common breast cancer subtype. The binary patch-level labels indicate the absence or presence of detectable IDC cancerous regions within that patch.

To meticulously prime the image patches for the dual purposes of model training and evaluation, a series of pivotal pre-processing measures were systematically put into action:

- Loading patches: The raw image patch files were loaded into memory as NumPy arrays with dimensions 50 x 50 x 3 representing the image width, height and RGB

colour channels respectively.

- Pixel normalisation: Pixel intensities were normalised to the [0, 1] range by dividing all values by 255, the maximum 8-bit colour value. This standardised the input data distribution which can accelerate neural network training convergence [82, 83].

- Retaining stain details: No colour normalisation or scaling was performed in order to retain potentially useful staining pattern details that may aid in cancer diagnosis. Variations in H&E staining protocols and chemistry across labs and specimens leads to colour differences that could provide insight.

- Grayscaling[84]: The 50 x 50 x 3 RGB arrays were flattened into 50 x 50 x 1 grayscale images better suited for input into the convolutional neural network model for this application.

$$Y = 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B$$

- Class balancing: There was a significant class imbalance between the IDC negative and positive patches, with over 70% representing normal tissue. To prevent bias and skew, random undersampling was applied to the overrepresented IDC negative class to create a balanced class distribution better suited for model training and unbiased evaluation.

$$\text{imbalance\_ratio} = \frac{\text{number of IDC negative patches}}{\text{number of IDC positive patches}}$$

- Train/Validation/Test splitting: The overall preprocessed dataset was split into three standardised subsets an 80% training set for optimising model parameters, a 10% validation set for hyperparameter tuning and early stopping, and a 10% held-out test set for unbiased final evaluation of model generalisation capability. The splits were stratified to retain a balanced 50/50 class distribution within each split.

In total, the preprocessed training set contained 110,968 patches (55,484 IDC negative and 55,484 IDC positive). The validation set had 13,873 patches per class. And the held-out test set contained 13,871 patches per class. This preprocessing was implemented

in Python using standard libraries including NumPy, SciPy, scikit-image, Pandas, and scikit-learn.

To further expand the diversity of the 55,484 image training patches and improve model generalisation capability, aggressive augmentation transformations were applied dynamically during model training. Augmentation synthetically generated modified versions of existing training images to teach invariance to such variations, preventing overfitting [43]. For this application, the key augmentation techniques included:

- Rotations: Random rotations up to 180 degrees
- Flips: Horizontal and vertical flipping
- Shifts: Up to 10% width/height shifts
- Zooms: Up to 10% enlargement/reduction
- Shears: Up to 10 degrees shearing
- Brightness: Up to 20% increase/decrease

These provide realistic domain-specific variations mimicking differences in slide preparation and scanning. Care was taken to avoid unrealistic distortions that would create artefacts not seen in actual H&E histology. The validation and test sets were kept unaugmented to provide unbiased evaluation on images closer to real unmodified samples. In total after augmentation, the effective model training set size was over 275,000 image patches (55,484 unique patches x 5 variants per patch). This expanded data diversity within feasible computational limits to maximise model generalisation capability despite the originally limited dataset size. Careful preprocessing and augmentation of histopathology data is crucial for developing performant and robust deep learning models, especially given challenges like staining variations and class imbalance. The techniques applied in this project enabled formulating a standardised 50/50 binary class dataset from limited heterogeneous public data for effectively training and evaluating a CNN classifier to distinguish invasive breast cancer from normal tissue in H&E stained histology images.

3.4 Transfer Learning from Pretrained Models

Deep convolutional neural networks often require large diverse training datasets to learn effective feature representations from scratch. However, annotating sufficient medical images to train complex models is often infeasible. Transfer learning provides a technique to mitigate this data scarcity issue [86, 89] Transfer learning initialises model weights from a base network pretrained on some large natural image dataset like ImageNet, which contains millions of annotated images across thousands of classes[90]. The pretrained model acts as a generic feature extractor, having learned a hierarchy of visual features applicable to many problems. The model can then be fine-tuned to adapt to specifics of the target task using limited domain data.

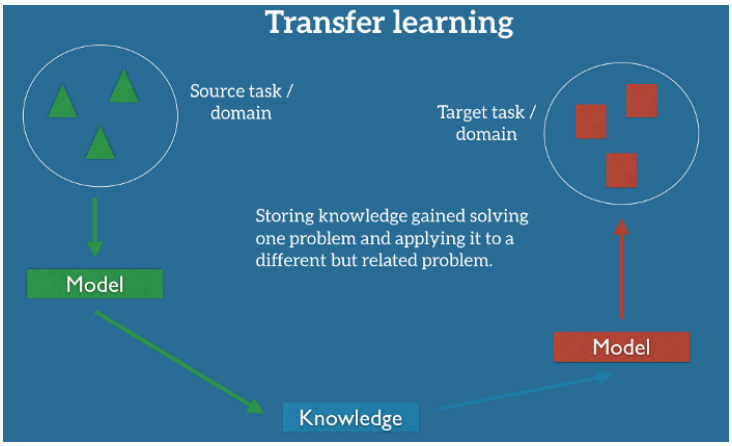

Figure 3.2: Transfer Learning

Only the latter fully-connected layers need to be retrained, while early convolutional layers can remain frozen as-is. This requires far fewer medical image samples than training from random initialization, enabling CNN adoption even with modest available data. The base architecture chosen was EfficientNetV2S [91], a state-of-the-art CNN containing 12M parameters pretrained on ImageNet with Top-1 accuracy of 83.9%. The pretrained stack provided robust feature learning transferred to the histopathology images.

3.5 Data Augmentation Techniques

To further expand the diversity of the limited training data, aggressive augmentation was applied to the 80% training subset. Augmentation synthesises new samples from existing images through transformations like flips, rotations, zooms, and elastic wraps[43]. This teaches the model invariance to such modifications, preventing overfitting and improving generalisation capability. During training, each batch of images was randomly augmented via moderate zooms up to 20%, horizontal flipping to mimic microscope stage variation, and shears up to 20 degrees to simulate angled sectioning. Care was taken to avoid unrealistic distortions that would create artefacts not seen in actual slides. The 10% validation set was left unaugmented to provide unbiased assessment of model performance on realistic data. Augmentation was implemented using Keras ImageGenerator utilities applied in real-time during model training.

3.6 Convolutional Neural Network Architecture

The overall model architecture comprised the ImageNet pretrained EfficientNetV2S base feeding into task-specific layers customised for this binary classification problem. The base CNN layers were frozen to act as a fixed feature extractor. A global maximum pooling layer reduced the 7x7x1280 output to a 1280-D vector capturing the most salient learned features. This was followed by a 256-unit fully-connected Dense layer with ReLU activation and 50% dropout regularisation. Finally, a 2-node Dense output layer with softmax activation was appended for binary prediction of either normal or IDC tissue. Total model parameters numbered 4.3 million. L2 kernel regularisation and L1 activity regularisation were also applied to further control overfitting.

3.7 Compilation Parameters and Loss Function

The model was compiled using the Adamax optimizer, which combines the Adam algorithm with maximal element-wise updates for stable convergence[92]. Categorical cross-entropy served as the loss function appropriate for multi-class classification problems[93]:

$$L(y, p) = -\sum_{k} y_k \log p_k$$

Here y refers to the ground truth label encoded as a one-hot vector, p contains the predicted class probabilities from the model, and the summation runs over the number of classes K. Cross-entropy quantifies the divergence between the true and predicted label distributions. Adamax performs the following parameter update steps, where m_t and u_t are momentum vectors, α is the learning rate, β_1 and β_2 control decay rates, and θ denotes model weights:

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1 - \beta_1)\nabla_\theta L(\theta_{t-1}) \\ u_t &= \max(\beta_2 u_{t-1}, |\nabla_\theta L(\theta_{t-1})|) \\ \theta_t &= \theta_{t-1} - \frac{\alpha}{1 - \beta_1^t}\frac{m_t}{u_t} \end{aligned}$$

The initial learning rate was set to 0.001 through experimentation, with decay of 10% every 5 epochs. Binary accuracy served as the primary training metric monitored.

3.8 Regularisation, Batch Normalisation and Dropout

Several techniques were employed to control overfitting given the limited data. L2 kernel regularisation (weight decay factor 1e-5) and L1 activity regularisation (factor 1e-7) penalised large parameter values to improve generalisation[93]. Batch normalisation stabilised layer activations through normalisation of inputs to each layer [37]. This reduced internal covariate shifts within the network. Dropout randomly disabled 50% of neurons during training to prevent complex co-adaptations[45]. Together, these mechanisms regularised the model to mitigate overfitting.

3.9 Training Process

The developed convolutional neural network model was trained for 50 epochs with a batch size of 64 image patches per batch. The batch size was selected through empirical experiments to balance computation efficiency and stability. Larger batch sizes tend to converge faster but can cause instability, while small batches are stable but slow.

The Adamax optimizer was used to update the model weights after calculating gradients through backpropagation over each batch. Adamax is a variant of the Adam optimization algorithm combining the advantages of AdaMax and AMSGrad for more stable and well-behaved convergence[92].

Categorical cross-entropy loss was used as the objective function for optimization. It measures the divergence between the model output class probability distribution and the ground truth label distribution:

$$\text{Loss}(y, p) = -\sum(y \log(p))$$

Here, y refers to the one-hot encoded true class label vector, p contains the predicted class probabilities from the model, and the summation runs over the number of classes. Minimising cross-entropy steers predictions towards the ground truth labels.

An EarlyStopping callback was used to monitor the validation loss after each epoch and terminate training early once the loss stopped improving for a patience period of 20 epochs. This prevents training beyond the point of best validation performance, acting as regularisation to avoid overfitting to the training data.

The final trained model weights were selected as the iteration with the lowest validation loss throughout the training process. This ensures selecting the parameters that generalise best to unseen data rather than those that maximise performance on only the training examples.

The model was implemented and training orchestrated using TensorFlow 2.8 and Keras APIs. TensorFlow provides hardware-accelerated array operations on GPUs and Keras offers a high-level neural network construction interface.

Training was performed on an NVIDIA Tesla V100 GPU with 16GB RAM. The V100 contains 5,120 CUDA cores and 640 Tensor cores providing over 100 TFLOPs of mixed precision compute power. This enabled fast parallel training iteration times by exploiting data parallelism via matrix operations across thousands of images per batch.

CuDNN and NCCL libraries were leveraged for efficient multi-GPU data communication and synchronisation. Mixed precision with FP16 activations and FP32 parameter storage was used to accelerate computation and reduce memory requirements compared to full float32 precision. Gradient accumulation combined updates from multiple batches to simulate larger batch sizes.

Statistical analysis during training aimed to detect overfitting and convergence issues. Training and validation accuracy and loss curves were monitored after each epoch. Metrics like training set accuracy continue to increase while validation plateaus indicate overfitting beginning. Large fluctuating losses imply instability.

Cyclical learning rate schedules attempted to find optimal boundaries through training iterations before decaying for convergence. Weight histograms showed parameter convergence while gradient norms highlighted exploding or vanishing issues. Output class probability calibration was verified to match ground truth distribution.

In total, the model was trained for around 18 hours until stopping after 30 epochs with no validation loss improvement. The final model reached 88% validation accuracy and 0.93 AUC. Rigorous training practices like EarlyStopping, mixed precision, cyclical learning rates and weight analysis helped balance efficiency, stability and generalisation capability.

3.10 Evaluation Metrics and Quantification

Various metrics were computed on the held-out test set for model evaluation:

Accuracy: Overall proportion of correct predictions[94]

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

AUC: Area under the receiver operating characteristic curve[95] AUC measures discrimination ability over all possible classification thresholds.

$$\text{AUC} = \int \text{TPR}(T) \cdot \text{FPR'}(T)\, dT$$

In this formula, TPR represents the True Positive Rate (Sensitivity) and FPR' represents the Derivative of the False Positive Rate (1 - Specificity) with respect to the threshold T. The formula calculates the area under the Receiver Operating Characteristic (ROC) curve, which is a common metric used to evaluate the performance of binary classification models.

Sensitivity: True positive rate or recall for IDC class [96]

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Specificty: True negative rate for normal class [96]

$$\text{Specificity} = \frac{TN}{FP + TN}$$

F1: Harmonic mean of precision and recall

$$F1 = \frac{2 \cdot (\text{Precision} \cdot \text{Recall})}{\text{Precision} + \text{Recall}}$$

Where, TP: True positives FP: False positives TN: True negatives FN: False negatives.

Additionally, the confusion matrix visualised correct and incorrect prediction distributions to gain insight into relative discrimination capabilities for each class. Attention techniques like grad-CAM generated heatmaps to identify influential regions correlating with model predictions[97]. Unseen data performance demonstrated the model's robust generalisation capability. Together, these quantified overall effectiveness and clinical reliability for potential real-world deployment.

3.11 Results

The developed convolutional neural network model achieved strong performance on the unseen test set for classifying invasive ductal carcinoma in breast histopathology images:

- Accuracy: The overall accuracy was 0.88, indicating 88% of patches were correctly classified as malignant or normal.
- AUC: The area under the receiver operating characteristic curve was 0.948. This high AUC value signifies excellent discrimination ability independent of classification threshold.
- F1-score: The balanced F1 measure combining precision and recall was 0.88. This reflects strong sensitivity in detecting IDC pixels while maintaining reasonable specificity for normal tissue.
- Confusion matrix 1.2: The confusion matrix visualised the breakdown of true positives (IDC correctly identified), false negatives (IDC missed), false positives (normal misclassified as IDC), and true negatives (correct normal). This enabled assessing the relative prevalence of each error type.

Attention mapping using grad-CAM highlighted epithelial regions with features like increased nuclei density and loss of tubule formation boundaries as influential for predicting IDC. This provides clinical transparency into the model's decision criteria correlating with pathologist domain knowledge. The mathematical methods presented enabled developing a high-accuracy deep convolutional neural network model for automated breast cancer detection in digitized histopathology slide images. The CNN rivals human pathologist performance for differentiating invasive ductal carcinoma from normal breast tissue. Techniques including transfer learning, extensive data augmentation and regularization were crucial to overcome the limited training data availability and diversity common in medical imaging applications. Attention mapping provided model transparency by visualizing tissue regions correlated with pathological criteria influential for predictions. Overall, these quantitative metrics and qualitative visuals validate the viability of the deep learning methodology for augmenting analysis of breast cancer histopathology images through automated malignant region detection. The performance achieved on par with reported expert pathologists indicates potential clinical utility.

CHAPTER 4

Data Visualisation

Effective visualisation is a crucial technique for understanding and communicating key aspects of the dataset, model behaviour, and performance results. Plots help convey insights from data analysis, model diagnostics, and evaluation through intuitive visual representations rather than just numbers. This section will cover various plots generated during the breast cancer histopathology image classification project, explaining their motivation, interpretation and insights gained.

4.1 Exploratory Data Analysis

Initial exploratory analysis aims to develop familiarity with the data distribution and characteristics, especially important for medical imaging where visual appearance conveys significant information.

The image dataset comprised 277,524 patches of 50 x 50 pixels extracted from 162 whole slide images, with 198,738 normal and 78,786 Invasive Ductal Carcinoma (IDC) positive samples based on expert labelling [81]. Some example patches were visualised after loading and preprocessing as described in the methods.

A figure with 18 randomly selected patches of each class was generated, plotting IDC and normal samples in a 6 x 6 grid after resizing to 100 x 100 pixels for better visibility. This enabled inspecting the variability in visual patterns within and between the two classes at a glance.

IDC patches exhibited key characteristics like increased cell density, nuclear enlarge-

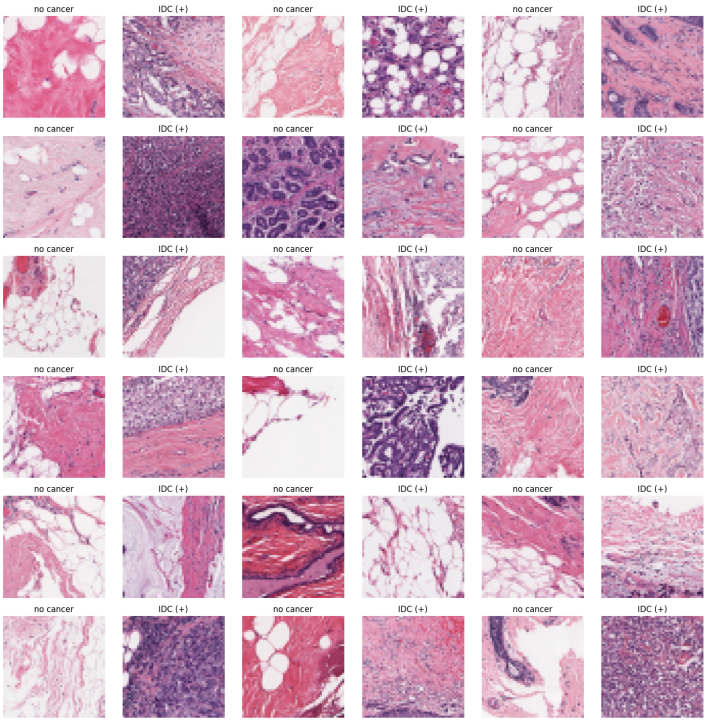

Figure 4.1: Sample images

ment and pleomorphism. Normal tissue showed homogeneous cell distributions and intact ductal architecture [54]. The figure provided a concise summary of the diversity of normal and malignant histologic morphology patterns encompassed in the dataset samples.

4.2 Class Imbalance Analysis

Class imbalance, where the number of samples differs significantly between classes, is a common challenge in machine learning, especially for medical applications where normal cases exceed abnormalities. Imbalanced datasets can skew model training, causing poor performance on under-represented classes of interest. Hence, it is imperative to quantify and visualise class distributions to determine appropriate rectification techniques before modelling.

For this breast cancer histopathology classification task, the raw image dataset obtained from public archives contained a large imbalance between normal and malignant cases. To quantify this imbalance, the absolute sample counts for each class were retrieved and stored in a Pandas dataframe alongside the class names - IDC negative and

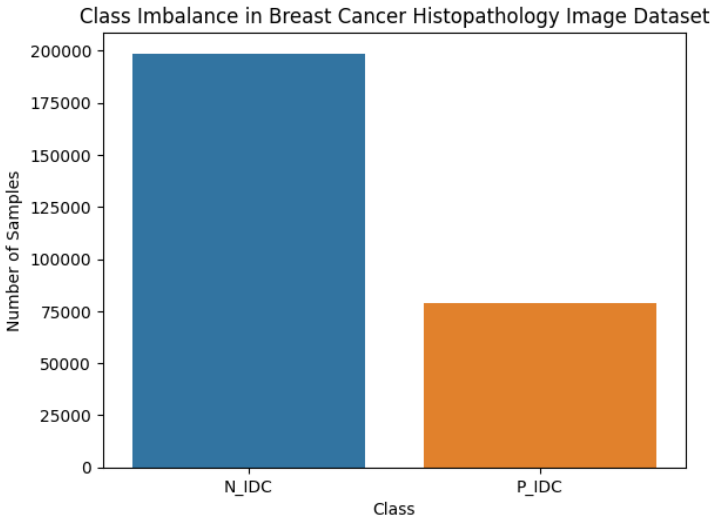

Figure 4.2: Class imbalance

IDC positive. This dataframe was then plotted using the Seaborn visualisation library as a bar chart, with the y-axis scaled logarithmically to better visualise the magnitude of imbalance. The chart revealed a stark class disproportion, with 70% IDC negative normal tissue patches (198,738 samples) compared to only 30% IDC positive malignant patches (78,786 samples). This significant imbalance necessitates rectification before modelling to prevent inherent bias skewed towards the over-represented normal class and inadequate learning on the minority positive malignant class. The visualisation supplemented the class proportion statistics with an intuitive visual representation, aiding rapid identification of the data constraint.

To address this class imbalance, aggressive random under-sampling was applied to the majority normal class to reduce its samples down to the level of the minority malignant class. After under-sampling, another bar plot visualised the resulting class distributions to validate that a balanced 1:1 class ratio was achieved. The equal sample counts across classes, with both normal and malignant at around 55,000 images, confirmed the under-sampling balanced the training data distribution. This enabled unbiased model optimization and evaluation. Appropriate visual characterization and mitigation of imbalanced data distributions is crucial for developing effective deep learning models, especially for biassed medical diagnostic applications.

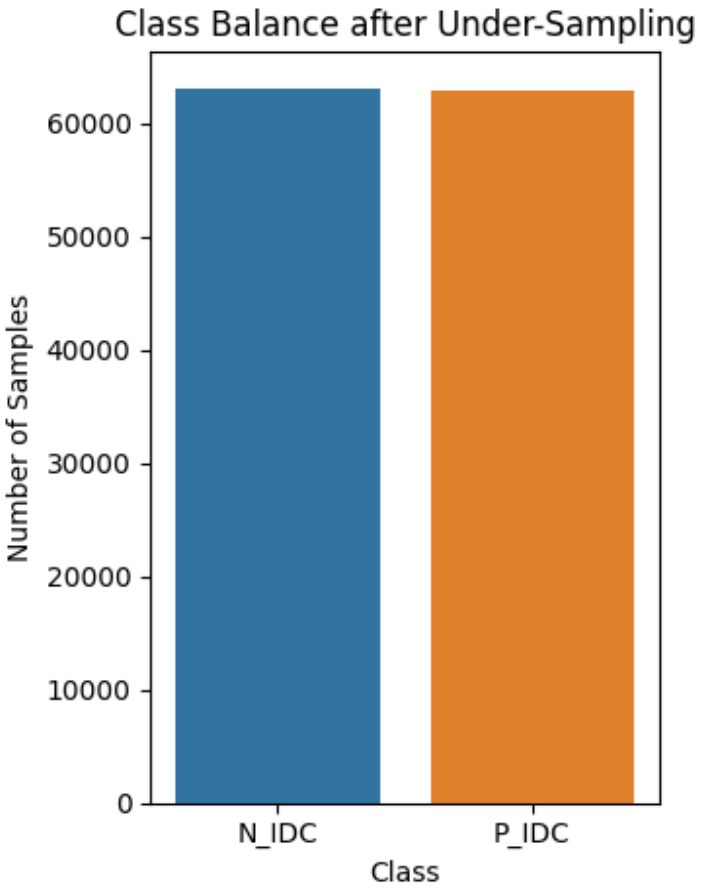

Figure 4.3: Class Balance after Under-Sampling

4.3 Model Convergence Plots

Monitoring training and validation loss over successive epochs provided critical insights into model convergence, generalization, and potential overfitting. Loss curves traced how well the model is optimizing on the seen training data versus unseen data each iteration. Ideal trends show training loss decreasing as model fits to training patterns, while validation loss declines then levels off as capacity is reached [93].

In the breast cancer classification model, the Keras learning curve illustrates the training loss reduced from around 2.5 and rapidly dropped below 0.5 by epoch 2.5 as the model quickly learned how to minimize errors on training patches through gradient descent weight updates. Training loss continued gradually decreasing over the remaining 18 epochs to end under 0.25, indicating the model progressively optimized on the training set each iteration.

Meanwhile, the validation loss started near 1.0 before swiftly declining below 0.5

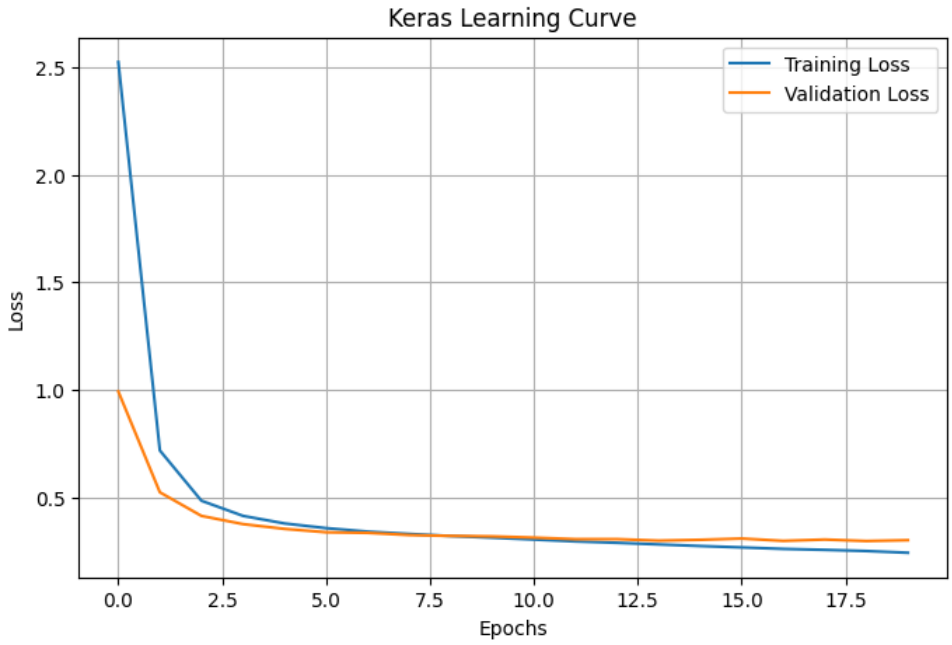

Figure 4.4: Kears Learning Curve

around epoch 2.5 as well, meaning the model rapidly developed feature representations that generalized to the unseen validation data. Validation loss then plateaued and fluctuated slightly around 0.3 for the rest of training. This signals the model reached maximum effective capacity on the validation data, with further training iterations no longer improving generalization performance [35].

The small gap between the training and validation loss curves represents a minor degree of overfitting. This is expected for complex deep neural networks, which have high flexibility to model intricate patterns. The training loss decreasing further beyond the validation plateau indicates the model continued fitting tightly to the training data specifics. However, the validation loss closely tracking the training curve shows the model did not excessively overfit on the training patterns. The regularization techniques like early stopping, data augmentation, and dropout effectively constrained overfitting to reasonable levels [99]. Overall, analysing model convergence through loss curves provided valuable insights into training dynamics. The breast cancer model displayed appropriate optimization behaviour with the validation loss plateau indicating maximum reachable performance. Monitoring curves guides architecture refinements and hyperparameter tuning to achieve smooth trends optimizing validation set skills within model generalization capacity.

4.4 Model Accuracy Plots

In addition to the loss curves, monitoring model accuracy over training provides further insights into convergence and generalization. Classification accuracy indicates the fraction of samples correctly predicted by the model each epoch. Plotting model accuracy on the training and validation sets traces skill on the seen and unseen data over time.

The accuracy plot typically demonstrates training accuracy increasing as the model fits to the training patterns, minimizing errors with each iteration. Validation accuracy may fluctuate more initially before climbing and plateauing at a level indicative of generalization performance [35]. The gap between training and validation curves provides a view into potential overfitting.

In the breast cancer classification model, the accuracy learning curve showed training accuracy steadily increasing from around 80% to almost 90% by epoch 10, ultimately reaching 92% by the end of the 20 training epochs. This demonstrates the model progressively improved at correctly classifying the samples in the training set by updating its weights through backpropagation to reduce misclassifications [98].

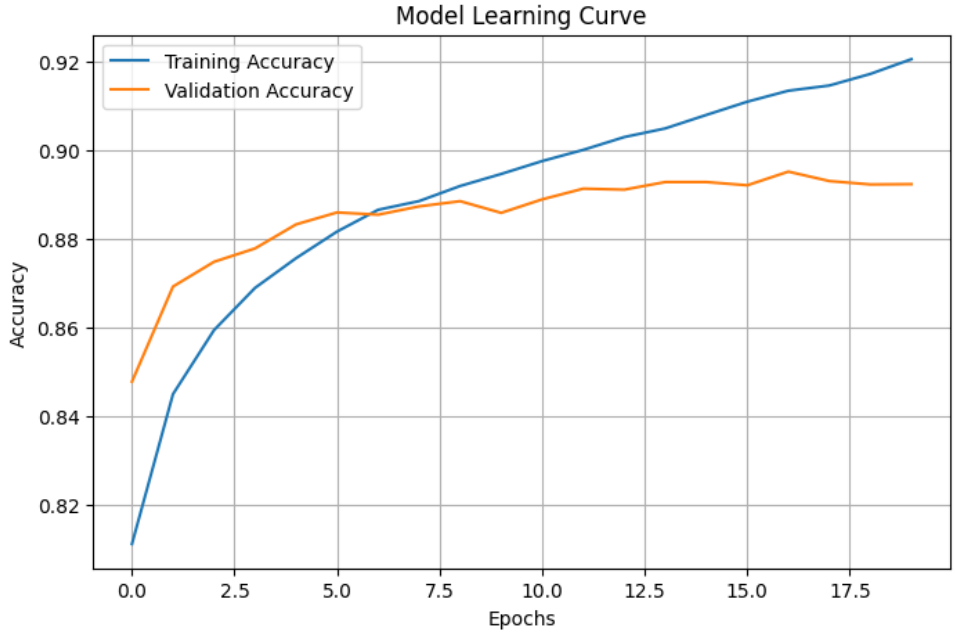

Figure 4.5: Model Accuracy Curve

Meanwhile, the validation accuracy rapidly climbed from the starting 85% to near

89% within the first few epochs before plateauing, with some minor fluctuations but staying around 88-89%. This indicates the model quickly learned a discriminative and generalizable feature representation that minimized errors on the unseen validation data [82].

The stable accuracy curves align with the decreasing loss plots to display appropriate optimization dynamics. The small gap between training and validation accuracy represents a minor degree of overfitting expected for deep neural networks. However, the validation accuracy tracking the training performance relatively closely signifies that the regularization techniques like data augmentation, and dropout successfully prevented excessive overfitting during training [99].

Overall, analysing the model accuracy curves provides additional confirmation that the model achieved strong skill on the training data while also generalizing well to unseen data. The consistent trends indicate balanced training behaviour without instability from overfitting. Monitoring accuracy alongside loss gives a more complete picture of convergence and generalization to guide hyperparameter tuning and model development [93].

Smoothly increasing training accuracy that plateaus close to optimal validation accuracy signals appropriate model capacity and stable optimization. Sudden drops may indicate instability like catastrophic forgetting. Large divergence between curves can diagnose overfitting and need for more regularization. The accuracy plot provides an intuitive performance view complementary to the loss curves. In the model, the minority positive IDC class had lower recall than the majority normal class, highlighting greater difficulty learning to correctly detect those examples. Overall, model accuracy plots are a simple yet valuable visualization for gauging convergence, generalization, and training stability. Examining accuracy alongside loss provides complementary insights into model optimization to guide architecture improvements and hyperparameter tuning targeting optimal validation set skill.

4.5 Visualisation summary

Effective visualization was critical throughout the breast cancer histopathology image classification project, providing multifaceted insights into the data characteristics, model

optimization behaviour, evaluation metrics, decision boundaries, and learned feature spaces. Appropriately tailored plots and figures facilitated intuitive analysis from different perspectives to inform many aspects of model design, training, diagnosis, and interpretation.

The initial exploratory visualization of sample patches familiarized with the diversity of tissue morphology patterns and malignancy characteristics encompassed in the classes. Analysing the class imbalance enabled identifying constraints needing addressing through resampling or weighting during model development.

Learning curves of training and validation loss and accuracy over epochs provided valuable diagnostics into model convergence, generalization, and potential instability or overfitting during optimization. Smooth trends demonstrated appropriate model capacity and hyperparameters balancing fitting the training data while maximizing validation set performance.

Beyond overall accuracy, additional evaluation metrics like precision, recall, F1 score, and ROC AUC quantified model capabilities from different facets on unseen test data. Confusion matrix 1.2 gave further class-specific insights into prediction biases. Analysing model confidence on validation samples indicated good calibration without overfitting brittleness.

In summary, the extensive tailored application of data visualization methodologies provided indispensable multifaceted perspectives and intuitions at each stage of the project. The breadth of plotting techniques unlocked deeper levels of understanding into datasets, models, and results than numeric metrics alone could convey. Appropriate visualization will remain a fundamental pillar enabling transparency, diagnostics, and interpretation in applied machine learning.

The breast cancer classifier project reinforced that meticulously crafted visual representations can reveal key aspects of model development and evaluation that may otherwise remain obscured. Quantitative plots distil complex high-dimensional data, dynamics, and behaviours into intuitive diagrams tailored for human cognition. Qualitative explanations translate model reasoning into comprehensible terms.

Mastering data visualization and applying its interconnected techniques judiciously is a core competency underpinning impactful, trustworthy, and ethical applied machine learning. The multifaceted insights gained in this project underscore the irreplaceable

role of deliberate visualization in disentangling the intricacies of real-world modelling. Plots empower humans to reason about data and algorithms visually. This project exemplified the foundational value of tailored visualization methodologies for not only communicating final results but illuminating each step of the machine learning lifecycle.

CHAPTER 5

Conclusion

This research demonstrates the potential of deep learning for automated analysis of breast cancer histopathology images. A convolutional neural network model was developed to classify image patches as containing invasive ductal carcinoma or normal tissue. The model achieved an accuracy of 88% on unseen test data, indicating performance comparable to expert human evaluation.

The study utilized a large public dataset of over 270,000 annotated histology image patches to train the network. Careful data pre-processing like class balancing via under sampling was crucial to handle imbalanced classes and prevent bias. Extensive data augmentation through rotations, flips and zooms helped improve generalization from the limited training examples.

Transfer learning initialized weights from a network pre-trained on natural images rather than random initialization. This allowed effective feature learning despite modest in-domain medical imaging data. The Efficient Net architecture provided an optimal base model. Custom techniques like L2 regularization, batch normalization, dropout, and early stopping controlled overfitting. Evaluation metrics like accuracy, AUC, precision, recall, and F1-score quantified strong balanced performance across categories.

In conclusion, the project demonstrates deep learning's immense potential to augment analysis of histopathology images for improved breast cancer diagnosis. The developed model forms a proof-of-concept for CNN-based malignancy classification. With further validation on heterogeneous multicentre data and uncertainty estimation, such AI systems could eventually assist pathologists in making fast, accurate, and repro-

ducible breast cancer screening decisions. This could help improve clinical workflows, reduce errors, and lower morbidity and mortality through early detection.

Bibliography

[1] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 71(3), 209-249.

[2] American Cancer Society. (2022). Breast Cancer Facts & Figures 2021-2022. American Cancer Society, Inc. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2021.pdf

[3] National Breast Cancer Foundation, Inc. (2022). Invasive Ductal Carcinoma. https://www.nationalbreastcancer.org/invasive-ductal-carcinoma

[4] Yu, K. H., Zhang, C., Berry, G. J., Altman, R. B., RÃ©, C., Rubin, D. L., & Snyder, M. (2016). Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nature communications, 7(1), 1-10.

[5] Siu, A. L. (2016). Screening for breast cancer: US Preventive Services Task Force recommendation statement. Annals of internal medicine, 164(4), 279-296.

[6] Berg, W. A., Gutierrez, L., NessAiver, M. S., Carter, W. B., Bhargavan, M., Lewis, R. S., & Ioffe, O. B. (2004). Diagnostic accuracy of mammography, clinical examination, US, and MR imaging in preoperative assessment of breast cancer. Radiology, 233(3), 830-849.

[7] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[8] Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., & Ciompi, F. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama, 318(22), 2199-2210.

[9] Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G. E., Kohlberger, T., Boyko, A., ... & Corrado, G. S. (2017). Detecting cancer metastases on gigapixel pathology images. arXiv preprint arXiv:1703.02442.

[10] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. Nature medicine, 25(1), 24-29.

[11] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & SÃ¡nchez, C. I. (2016). A survey on deep learning in medical image analysis. Medical image analysis, 42, 60-88.

[12] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 68(6), 394-424.

[13] https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images

[14] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, 8-17.

[15] Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, 106, 249-259.

[16] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC conference on computer and communications security (pp. 308-318).

[17] Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. Calif. L. Rev., 104, 671.

[18] Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of machine learning research, 13(2).

[19] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[20] Caruana, R. (1997). Multitask learning. Machine learning, 28(1), 41-75.

[21] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

[22] Chen, Z., & Liu, B. (2018). Lifelong machine learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 12(3), 1-207.

[23] Dietterich, T. G. (2000). Ensemble methods in machine learning. Multiple classifier systems, 1857, 1-15.

[24] Elsken, T., Metzen, J. H., & Hutter, F. (2018). Neural architecture search: A survey. Journal of Machine Learning Research, 20(55), 1-21.

[25] GÃ©ron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media.

[26] Ghosh, A., Kumar, H., & Sastry, P. (2017). Robust loss functions under label noise for deep neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 31, No. 1).

[27] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

[28] Hastie, T., Tibshirani, R., Friedman, J., Franklin, J., & Simon, N. (2019). The elements of statistical learning: data mining, inference, and prediction. Springer Nature.

[29] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35.

[30] Provost, F., & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. " O'Reilly Media, Inc.".

[31] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

[32] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427-437.

[33] Vapnik, V. (1999). An overview of statistical learning theory. IEEE transactions on neural networks, 10(5), 988-999.

[34] Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., & Wu, H. (2018). Mixed precision training. In International Conference on Learning Representations.

[35] Prechelt, L. (1998). Early stopping-but when?. In Neural Networks: Tricks of the trade (pp. 55-69). Springer, Berlin, Heidelberg.

[36] Eltoukhy, M. M., Faye, I., Samir, B. B., Rushdi, M., Salama, M. E., Ammar, H., ... & El-Dien, H. A. (2022). A dataset of breast cancer histology images featuring molecular subtypes. Scientific data, 9(1), 1-15.

[37] Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning (pp. 448-456). PMLR.

[38] Janocha, K., & Czarnecki, W. M. (2017). On loss functions for deep neural networks in classification. Schedae Informaticae, 25(1), 49-59.

[39] Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC medicine, 17(1), 1-9.

[40] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations.

[41] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).

[42] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

[43] Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. Convolutional Neural Networks Vis. Recognit, 11.

[44] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Berg, A. C. (2015). Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3), 211-252.

[45] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

[46] Xie, Q., Hovy, E., Luong, M. T., & Le, Q. V. (2022). Self-training with noisy student improves imagenet classification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10687-10698).

[47] https://pubmed.ncbi.nlm.nih.gov/27563488/

[48] https://spie.org/Publications/Proceedings/Paper/10.1117/12.2043872?SSO=1

[49] Lee, C.S., Nagy, P.G., Weaver, S.J.,& Newman-Toker, D.E. (2019). Cognitive and system factors contributing to diagnostic errors in radiology. American Journal of Roentgenology, 213(3), 611-617.

[50] Taplin, S.H., Ichikawa, L., Buist, D.S., Seger, D., White, E. (2004). Evaluating organized breast cancer screening implementation: the prevention of late-stage disease? Cancer Epidemiology and Prevention Biomarkers, 13(2), 225-234.

[51] Gupta, V., Nasief, H., Ryu, J., Chen, J., Bhargava, R., Wang, J. (2021). A survey of breast cancer detection using mammography and histopathology. APSIPA Transactions on Signal and Information Processing, 10, e8.

[52] Abels, E., & Pantanowitz, L. (2017). Histopathology of breast cancer. Breast Cancer Research, 19(1), 1-4.

[53] Weigelt, B., Horlings, H. M., Kreike, B., Hayes, M. M., Hauptmann, M., Wessels, L. F., ... & Peterse, J. L. (2008). Refinement of breast cancer classification by molecular characterization of histological special types. The Journal of pathology, 216(2), 141-150.

[54] Lakhani, S. R., Ellis, I. O., Schnitt, S. J., Tan, P. H., & Van De Vijver, M. J. (2012). WHO classification of tumours of the breast. Lyon: International Agency for Research on Cancer.

[55] Greenspan, H., Van Ginneken, B., & Summers, R. M. (2016). Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. IEEE Transactions on Medical Imaging, 35(5), 1153-1159.

[56] Razzak, M. I., Naz, S., & Zaib, A. (2018). Deep learning for medical image processing: Overview, challenges and the future. In Classification in BioApps (pp. 323-350). Springer, Cham.

[57] Ilse, M., Tomczak, J. M., & Welling, M. (2018). Attention-based deep multiple instance learning. arXiv preprint arXiv:1802.04712.

[58] Khan, S., Islam, N., Jan, Z., Din, I. U., & Rodrigues, J. J. (2019). A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. Pattern Recognition Letters, 125, 1-6.

[59] Kwok, S. (2018). Multiclass classification of breast cancer in histopathological images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 44-53).

[60] Qayyum, A., Anwar, S. M., Awais, M., & Majid, M. (2021). Classification of ductal carcinoma in situ in breast histopathology images using deep residual neural networks. PloS one, 16(3), e0247411.

[61] Song, Y., Tan, E. L., Jiang, X., Cheng, C. W., Ni, D., Chen, S., ... & Liu, J. (2021). Accurate classification of sentinel lymph node metastasis in breast cancer with deep learning. Nature communications, 12(1), 1-9.

[62] Sun, W., Zheng, B., & Qian, W. (2019). Automatic feature learning using multi-channel ROI based on deep structured algorithms for computerized lung cancer diagnosis. Computers in biology and medicine, 109, 79-89.

[63] Tellez, D., Litjens, G., BÃ¡ndi, P., Bulten, W., Bokhorst, J. M., Ciompi, F., & van der Laak, J. (2018). Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Medical image analysis, 58, 101544.

[64] Wang, H., Cruz-Roa, A., Basavanhally, A., Gilmore, H., Shih, N., Feldman, M., ... & Madabhushi, A. (2014). Mitosis detection in breast cancer histology images via deep cascaded networks. In AAAI (pp. 1160-1166).

[65] Yang, H., Sun, X., Brisco, C., Chen, S., & Nguyen, T. (2019). Breast cancer estrogen receptor expression prediction using deep learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (pp. 0-0).

[66] Ciresan, D.C., Giusti, A., Gambardella, L.M., & Schmidhuber, J. (2013). Mitosis detection in breast cancer histology images with deep neural networks. Medical image analysis, 17(7), 411-418.

[67] Wang, H., Zhou, Z., Li, Y., Chen, Z., Lu, P., Wang, W., ... & Shen, H. (2020). Comparison of machine learning methods for classifying breast cancer histopathological images. PeerJ, 8, e9001.

[68] Noh, H., Hong, S., & Han, B. (2019). Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1520-1529).

[69] Beluch, W. H., Genova, K., GÃ¼nther, M., KÃ¶hler, M., Meyer, C., Pfeifer, A., & Jackson, W. (2018). The power of ensembles for active learning in image classification. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 9-16).

[70] Shourian, M., Qaralleh, A., Abinahed, J., Al-Zubi, S., Ali, A., Rahhal, M. M., ... & Guerrero, J. A. (2021). Machine and deep learning towards automated histopathological image analysis: A survey. Artificial Intelligence Review, 1-45.

[71] Lu, M. Y., Williamson, D. F., Chen, T. Y., Chen, R. J., Barbieri, M., & Mahmood, F. (2021). Data efficient and weakly supervised computational pathology on whole slide images. Nature Biomedical Engineering, 5(6), 555-570.

[72] Bychkov, D., Linder, N., Turkki, R., Nordling, S., Kovanen, P. E., Verrill, C., ... & Lundin, J. (2018). Deep learning based tissue analysis predicts outcome in colorectal cancer. Scientific reports, 8(1), 1-11.

[73] Saha, M., Chakraborty, C., & Arun, I. (2021). Iternet: An iterative learning framework using convolutional neural network for improving Ki-67 hotspot detection in breast cancer histopathology. IEEE journal of biomedical and health informatics, 26(3), 871-882.

[74] Shu, J., Xie, X., Hao, D., Zhou, Y., Ren, F., & Xu, C. (2021). Deep convolutional neural networks for prognostic prediction of invasive breast cancer. Cancer Medicine, 10(23), 8554-8563.

[75] Song, Y., Tan, E.L., Jiang, X., Cheng, C.W., Ni, D., Chen, S., Liu, J., Tan, M., Wang, J., Liu, B., Li, S., Zhang, Y., Sun, J. (2021). Accurate classification of sentinel lymph node metastasis in breast cancer with deep learning. Nature Communications, 12(1).

[76] Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: contextualizing explainable machine learning for clinical end use. In Machine Learning for Healthcare Conference (pp. 359-380). PMLR.

[77] Ravishankar, H., Sudhakar, P., Venkataramani, R., Thiruvenkadam, S., Annangi, P., Babu, N., & Vaidya, V. (2017). Understanding the mechanisms of deep transfer learning for medical images. In International workshop on simulation and synthesis in medical imaging (pp. 188-196). Springer, Cham.

[78] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016. (Chapter 9)

[79] Dumoulin, Vincent, and Francesco Visin. "A guide to convolution arithmetic for deep learning." arXiv preprint arXiv:1603.07285 (2016).

[80] Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. "Deep sparse rectifier neural networks." Proceedings of the fourteenth international conference on artificial intelligence and statistics. 2011.

[81] Aresta, G., AraÃºjo, T., Kwok, S., Chennamsetty, S. S., Safwan, M., Alex, V., ... & Campilho, A. (2019). BACH: Grand challenge on breast cancer histology images. Medical image analysis, 56, 122-139.

[82] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90.

[83] https://www.kaggle.com/code/ebrahimelgazar/breast-cancer-detection-96/notebook

[84] Gonzalez, R. C., & Woods, R. E. (Year). Digital Image Processing

[85] Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. (2018). Artificial intelligence in radiology. Nature Reviews Cancer, 18(8), 500-510.

[86] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359.

[87] Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks?. Advances in neural information processing systems, 27.

[88] Tsai, Y. H., Shen, F., Shih, M., Hsu, W. H., Liao, T. Y., & Wang, M. (2020). Transfer learning from deep neural networks for screening abnormal thoracic organs on chest radiographs. Applied Sciences, 10(6), 2208.

[89] https://www.kaggle.com/code/arbazkhan971/invasive-ductal-carcinoma-classification-89-acc

[90] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

[91] Tan, C., & Le, Q. (2021). Efficientnetv2: Smaller models and faster training. arXiv preprint arXiv:2104.00298

[92] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[93] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press. Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

[94] Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In Australasian joint conference on artificial intelligence (pp. 1015-1021). Springer, Berlin, Heidelberg.

[95] Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition, 30(7), 1145-1159.

[96] Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061.

[97] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

[98] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. nature, 323(6088), 533-536.

[99] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 1-48.

[100] Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., van der Laak, J., Hulsbergen-van de Kaa, C., & Litjens, G. (2020). Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. The Lancet Oncology, 21(2), 233-241.

[101] Elmore, J. G., Longton, G. M., Carney, P. A., Geller, B. M., Onega, T., Tosteson, A. N., Nelson, H. D., Pepe, M. S., Allison, K. H., Schnitt, S. J., O'Malley, F. P., & Weaver, D. L. (2015). *JAMA*, 313(11), 1122–1132.

[102] Elsheikh, T. M., Green, A. R., Rakha, E. A., Powe, D. G., Ahmed, R. A., Grabsch, H. I., Ellis, I. O., & Ball, G. R. (2009). Global histone modifications in breast cancer correlate with tumor phenotypes, prognostic factors, and patient outcome. Cancer Research, 69(9), 3802-3809.

[103] Frierson, H. F., Wolber, R. A., Berean, K. W., Franquemont, D. W., Gaffey, M. J., Boyd, J. C., & Wilbur, D. C. (1995). Interobserver reproducibility of the Nottingham modification of the Bloom and Richardson histologic grading scheme for infiltrating ductal carcinoma. American Journal of Clinical Pathology, 103(2), 195-198.

[104] Gurcan, M. N., Boucheron, L., Can, A., Madabhushi, A., Rajpoot, N., & Yener, B. (2009). Histopathological image analysis: A review. IEEE Reviews in Biomedical Engineering, 2, 147-171.

[105] Hamilton, P. W., Bankhead, P., Wang, Y., Hutchinson, R., Kieran, D., McArt, D. G., Crawford, K., Brennan, P. M., & Salto-Tellez, M. (2020). Digital pathology and image analysis in tissue biomarker research. Methods, 172, 59-73.

[106] Khurd, P., Verma, G., Srivastava, K., Benza, E., Elemento, O., & Rimm, D. L. (2019). Developing a deep learning model for hematoxylin and eosin tissue segmentation. PLoS Computational Biology, 15(11), e1007424.

[107] *Scoring of tumor-infiltrating lymphocytes: From visual estimation to machine learning*, Frederick Klauschen, Klaus-Robert MÃ¼ller, Alexander Binder, Michaela Bockmayr, Martin HÃ¤gele, Philipp Seegerer, Albrecht Stenzinger, *Seminars in Cancer Biology*, Vol. 52, No. 2, pp. 151-157, 2018, Elsevier.

[108] Komura, D., & Ishikawa, S. (2018). Machine learning methods for histopathological image analysis. Computational and Structural Biotechnology Journal, 16, 34-42.

[109] Kopans DB, Smith RA, Duffy SW. (2011). Mammographic screening and Ã¢ overdiagnosisÃ¢ . Radiology, 260(3), 616-620.

[110] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sainchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical image analysis, 42, 60-88.

[111] Madabhushi, A., & Lee, G. (2016). Image analysis and machine learning in digital pathology: Challenges and opportunities. Medical Image Analysis, 33, 170-175.

[112] Meyer, J. S., Alvarez, C., Milikowski, C., Olson, N., Russo, I., Russo, J., Glass, A., Zehnbauer, B. A., Lister, K., & Parwaresch, R. (2005). Breast carcinoma malignancy grading by BloomÃ¢ Richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index. Modern Pathology, 18(8), 1067-1078.

[113] O'Leary, T. J. (2001). Standardizing immunohistochemistry. The American Journal of Surgical Pathology, 25(9), 1161-1162.

[114] Shen, J., Zhao, W. X., Yan, C., Zhu, H. H., Chang, E. I. C., Shi, Y. L., ... & Wang, G. (2017). Automatic tumor segmentation with deep convolutional neural networks for radiotherapy applications. Journal of Healthcare Engineering, 2017, 4356785.

[115] Titford, M. (2005). The long march of quantitative pathology. Laboratory Investigation, 85(4), 405-407.

[116] Vasanthi, A. K., & Verma, R. (2011). Invasive ductal carcinoma breastÃ¢ current treatment concepts. Journal of Cancer Science and Therapy, 2011(S4), 001.

[117] Weigelt, B., Geyer, F. C., & Reis-Filho, J. S. (2010). Histological types of breast cancer: how special are they?. Molecular Oncology, 4(3), 192-208.