**Using histopathological images to develop machine learning models for early detection and diagnosis of breast cancer**

**Acknowledgements**

**Abstract**

**Table of Contents**

[1. Introduction 6](#_Toc185022899)

[1.1 Overview 6](#_Toc185022900)

[1.2 Background 6](#_Toc185022901)

[1.3 Rationale 6](#_Toc185022902)

[1.4 Problem Statement 7](#_Toc185022903)

[1.5 Research Aim 7](#_Toc185022904)

[1.6 Research Objectives 7](#_Toc185022905)

[1.7 Research Questions 8](#_Toc185022906)

[1.8 Research significances 8](#_Toc185022907)

[1.9 Research structure 9](#_Toc185022908)

[1.10 Summary 9](#_Toc185022909)

[Chapter 2: Literature Review 11](#_Toc185022910)

[2.1 Overview 11](#_Toc185022911)

[2.2 Advancements in Histopathological Image Analysis for Breast Cancer Detection 11](#_Toc185022912)

[2.3 Application of Transfer Learning in Histopathological Image Analysis 12](#_Toc185022913)

[2.4 Integration of Multi-modal Data for Improved Diagnostic Precision 13](#_Toc185022914)

[2.4 Theoretical Underpinnings 14](#_Toc185022915)

[2.5 Literature Gap 15](#_Toc185022916)

[2.6 Chapter Summary 15](#_Toc185022917)

[Chapter 3: Research Methodology 16](#_Toc185022918)

[3.1 Overview 16](#_Toc185022919)

[3.2 Research Design 16](#_Toc185022920)

[3.3 Data Source and Description 16](#_Toc185022921)

[3.4 Data Pre-processing 17](#_Toc185022922)

[3.4.1 Directory Structure and Labelling 17](#_Toc185022923)

[3.4.2 Image Resizing and Normalization 17](#_Toc185022924)

[3.4.3 Data Augmentation 17](#_Toc185022925)

[3.5 Model Selection: VGG16 18](#_Toc185022926)

[3.5.1 Rationale for Transfer Learning 18](#_Toc185022927)

[3.5.2 Model Architecture 18](#_Toc185022928)

[3.6 Training Procedure 18](#_Toc185022929)

[3.6.1 Data Splits 18](#_Toc185022930)

[3.6.2 Hyper parameter Settings 18](#_Toc185022931)

[3.6.3 Early Stopping and Regularization 19](#_Toc185022932)

[3.7 Evaluation Metrics and Techniques 19](#_Toc185022933)

[3.7.1 Primary Metrics 19](#_Toc185022934)

[3.7.2 Secondary Metrics 19](#_Toc185022935)

[3.8 Hardware and Software 19](#_Toc185022936)

[3.8.1 Computing Environment 19](#_Toc185022937)

[3.8.2 Libraries and Frameworks 19](#_Toc185022938)

[3.9 Ethical and Reproducibility Considerations 20](#_Toc185022939)

[3.9.1 Ethical Use of Medical Data 20](#_Toc185022940)

[3.9.2 Reproducibility 20](#_Toc185022941)

[3.10 Summary 20](#_Toc185022942)

[References 21](#_Toc185022943)

# 1. Introduction

## 1.1 Overview

The introduction begins from the necessity and importance of the early-stage breast cancer detection and discusses the role of machine learning approach. This section presents the research problem, goals, and questions; it stresses CNNs’ application and the BreakHis dataset. This chapter justifies the need for the current study and gives an overview of the organization of the dissertation.

## 1.2 Background

Breast cancer has been identified as the second biggest cause of cancer deaths among women globally. Detecting the disease at an initial stage and making correct and timely diagnosis is vital for enhancing survival and therapeutic outlook. Histopathological examination of breast tissue is still the gold standard in diagnosis; however, often interpretation is performed manually and may be inaccurate (Zeiser *et al.* 2021). Barely a few years ago, there has been considerable discussions about using Artificial Intelligence (AI) and Machine Learning (ML) in the diagnosis of certain ailments by enhancing on their precision, reliability, and speed.

Convolutional Neural Networks (CNNs) which are a part of the larger deep learning model family constitute one of the most potent hybrid models for image classification especially in medical imaging. CNN can learn and extract features from huge amount of data of labeled images for extracting features, it is ‘automatic’ and there is no need for ‘feature engineering’. This dissertation aims at designing a new CNN based model for categorizing histopathological images of breast tumor tissue using the dataset BreakHis.

## 1.3 Rationale

Breast cancer remains to be a common cause of death among women globally proving the need for proper diagnosis of the disease in its early stages. Conventional techniques, cropping to image analysis and manual reporting of HPE results, are incommodious and bear high inter-observer variability. Conventional Machine learning (ML), particularly Convolutional Neural Networks (CNNs) is an ideal solution in that it provides automated image analysis and an accurate classification outcome. Thus, this research intends to look at the generalization problem in image magnification and variations using the BreakHis dataset (Benhammou *et al.* 2020). Another interesting feature of the dataset is that several magnification levels have been included whereby boosting the model’s robustness and feature discernment. Building an ML diagnostic tool can enhance pathologists’ clinical practice and improve efficiency and accuracy of their work. This study is important to promote AI-driven innovations of healthcare as well as increase favorable results of breast cancer treatment.

## 1.4 Problem Statement

Breast cancer is still a global social health issue with millions of new patients diagnosed yearly. Thus, diagnosis at a primary stage and with a high degree of probability is vital for designing possible outcomes and enhance the near existence percentage. The identification of histopathological images with traditional techniques and visual interpretation are time-consuming and explorative with inter-observer variability. The difficulties outlined above make a push for efficient and accurate diagnostic work (Heng *et al.* 2021).

Machine learning, especially Convolutional Neural Networks (CNNs) could be the best approach for classifying histopathological image with the use of the following benefits. However, a number of limitations remain, for instance images acquired at various magnifications have variable quality and sometimes the tissue architecture is heterogeneous and the staining may vary. Current models may not necessarily generalize well across different data sets whereby they may operate with high efficiency. To address these issues in this research, an improved CNN-based model is formulated based on the BreakHis dataset. The goal of the study is to improve the classification performance using fewer diagnostic discrepancies that will aid in identifying key features important in accurate differentiation of breast cancer.

## 1.5 Research Aim

The aim of this study is to establish a baseline Convolutional Neural Network (CNN) model for the classification of breast tumor histopathological images into benign and malignant categories, with the help of BreakHis dataset to improve diagnostic yield.

## 1.6 Research Objectives

The objectives of this study are:

* To build and test machine learning algorithms for the categorization of breast cancer from histopathological images.
* To determine the essential characteristics employed in histopathological images for proper cancer diagnosis.
* To test various algorithms of machine learning for distinguishing between two classes namely, benign and malignant types of breast cancerous tumor.

## 1.7 Research Questions

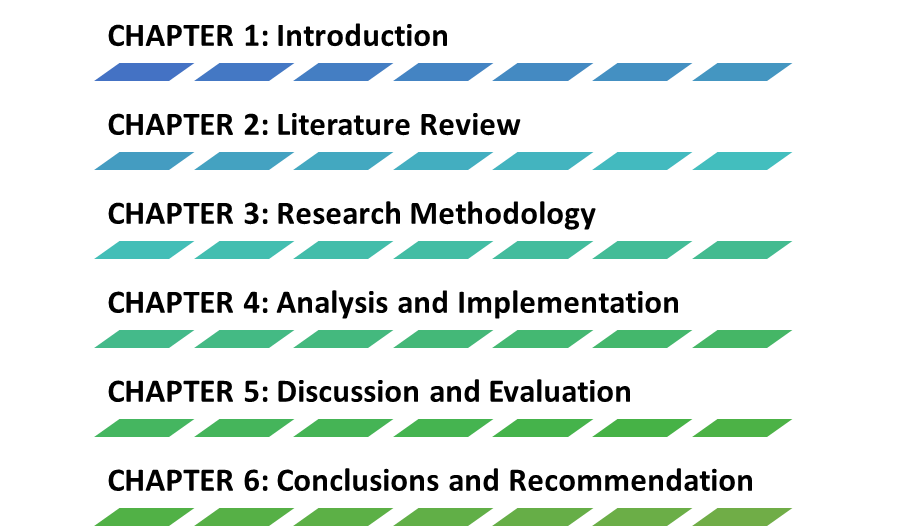
1. How effectively can machine learning models classify breast cancer using histopathological images from the BreakHis dataset?
2. What are the most significant features in histopathological images that contribute to accurate breast cancer detection and diagnosis?

## 1.8 Research significances

This study has great potential for further development for the use of machine learning for diagnosis of illnesses especially for early detection and staging of breast cancer. To this end, using Convolutional Neural Networks (CNNs), the research seeks to enhance diagnostic precision and coherency with ambitions toward eradicating restrictions of conventional visual interpretation of histopathological images. Tumor classification is tremendously important for timely intervention and treatment; failure to classify may lead to high mortality rates (Conte *et al.* 2022).

It also provides useful information for AI engineering in healthcare; for example, issues like image variability and model genera lization are discussed in the study. Analyzing the BreakHis dataset, the work introduces new approaches based on transfer learning and feature extraction to identify specific characteristics that may indicate cancer presence. In addition to technical contributions, the findings may be helpful to pathologists to guide clinical activities, decreasing the extensive or inaccurate identification of disease. Finally, this research creates the foundation for the discovery of dependable artificial intelligence diagnostic technique within the sphere of oncology.

## 1.9 Research structure



#### Figure 1.1 Outline of research

(Source: Self-created)

This figure shows the division of a dissertation into six principle chapters and each chapter has colored progress bar to demonstrate its significance and sequence of the dissertation. Chapter 1, “Introduction,” gives an idea on the background, problem, objectives and the importance of the study. Chapter 2 has been titled “Literature Review,” where the author examines prior work to determine the research’s place in previous literature. Chapter IV presents an overview of Research Methodology; this discusses the approach, techniques and strategies for data collection and analysis. Chapter 4 ‘’Analysis and Implementation’’ brings out the issue of presenting and interpreting results that were analyzed from the applied methodology. Chapter 5, whose title is ‘Discussion and Evaluation,’ provides a critical discussion of the results given in answer to the research questions and the existing literature. Last of all the Chapter 7 “Conclusions and Recommendations” presents research findings, limitations and suggestions for further research. The progressive shading also makes the reader appreciate the step-by-step dependency and harmony of each chapter to meet the envisaged objectives of the dissertation.

## 1.10 Summary

The introduction chapter contains the explanation of the problem and relevance of early detection of breast cancer as well as the relevance of machine learning algorithms in rising the accuracy of diagnosis. It discusses the research problem, goals and questions and the focus on CNN and the BreakHis dataset. The current chapter sets the tone of the study and gives the reader a preview of the contents of the book.

# Chapter 2: Literature Review

## 2.1 Overview

Breast cancer diagnosis using a machine learning approach utilizing histopathological images involves methods such as CNN and transfer learning for automatic detection and classification of tumours. These models help improve diagnostic performance but present several difficulties because multimodal data sources are insufficient and are required to improve performance and generalization.

## 2.2 Advancements in Histopathological Image Analysis for Breast Cancer Detection

Over the last few years, the analysis of histopathological images has grown massively, especially when it comes to screening breast cancer. Scientific advances in the applications of ML, AI, and various image-processing methods have accelerated the improvement in diagnosis. Biopsy Histopathological images, which are actual photographic images of affected tissues, offer essential information on the cells, tissue architecture, and prognosis of cancerous growth. Heretofore, such photos were often analyzed with the help of pathologists, but due to present-day innovations in computation, there are possibilities for improving the diagnosis of images. According to Srinidhi *et al.* (2021), A significant development in histopathological image analysis is using deep learning techniques, especially convolutional neural networks (CNNs). CNNs are intended to identify patterns in image data naturally and, as such, help analyze tissue structures in histopathological slides. These algorithms can identify the cancer cells, differentiate different kinds of tumours, and even estimate the malignancy degrees of tumours by assessing the sample’s characteristics depending on their morphology, including the cells’ shape, texture, and density.

Compared with other image processing methods, deep learning-based methods do not rely on owners to extract features since they directly extract features from raw image data in multiple layers. Thus, it leads to fewer interpretation errors and more specificity than human interpretation. Notably, incorporating transfer learning has further enhanced the existing deep learning models, including when results arise from a small amount of data. This means that transfer learning uses several priori-trained models in service of large databases, like Image Net, and trains them for particular applications, such as breast cancer detection. As stated by Ma *et al.* (2020), this approach has been instrumental in medical markup, as collecting large amounts of marked data is often costly and time-consuming. It provides an opportunity to learn from the data sets of greater volume and, therefore, have a higher probability of achieving relevant diagnostic accuracies. Another significant advancement is integrating machine vision as a quantitative image analysis supplement to deep learning.

Biological processes such as texture and morphometric analysis facilitate improved assessment of tissue heterogeneity, which is vital for proper tumour grading and staging. Fundamentally, these approaches enable one to be more refined when looking at the tumour attributes that include cell size, nuclear shape, and the extent of differentiation that otherwise would not exist or would be pretty obscure at first glance. As per the view of Foran *et al.* (2022), It enhances the information that pathologists feed the clinicians and researchers to society, giving a much richer picture of tumour behaviour. Besides staining methods, new technologies in multi-projection and multi-spectral organs and multi-modal imaging have promoted breast cancer detection. For instance, detecting molecular markers for cancer progression is realised by employing immunohistochemistry (IHC) and fluorescence microscopy, where the two form a composite system. This makes it possible to characterise tumour types more accurately, especially in the context of molecular stratification of cancer, where treatment strategies depend on the molecular characteristics of the patient’s tumour.

The use of automated diagnostic systems based on AI has also been indicated to be capable of giving pathologists a break. Such systems have several applications, including tumour detection, segmentation, and classification, to name but a few, and they are far better than most manual analyses. This paper examined various approaches to promulgate incorporating AI-assisted tools into a hospital's operation to increase examination standardization, decrease variability, and alleviate pathologists’ burden in centres with a high patient volume.

## 2.3 Application of Transfer Learning in Histopathological Image Analysis

It is also an essential part of diagnostic and prognostic disease management, specifically in oncology, which involves the examination of tissue samples under a microscope. Such images have rather complex features and patterns that can only be interpreted with sophisticated methods. Indeed, the conventional techniques in which the analysis relied on manual rather than automated analysis were time-consuming and beset with the risk of human error and inconsistencies that called for establishing optimum analyzing systems. According to Kora *et al.* (2022), transfer learning has recently been considered one of the practical approaches for many applications of medical image analysis, especially for histopathological image classification and segmentation. Transfer learning can be described as a machine learning methodology whereby the introduction learning model, which was trained on a large and diverse set, is altered to learn another set of tasks in a scope related to the former.

This approach is beneficial when acquiring a large dataset of labelled images is difficult and costly, as in the medical imaging domain. The basic idea of transfer learning is that unlike training a model from scratch, we fine-tune a base pre-trained model, such as one that has been trained using the ImageNet dataset, to extract general features of histopathological images and then train it for the particular problem using a small set of histopathological images.  As stated by Hamida *et al.* (2021), A significant problem in histopathological image analysis is the lack of annotated images, which is imperative for training deep neural networks and evaluating performance. Transfer learning solves this problem by training models on many datasets, perhaps in other domains, and then applying the knowledge learned to relatively small specialized datasets. Hence it can significantly improve the performance of models even when small histopathological data sets are available.

By training deep learning models from scratch, the computational costs and time required are notably high, especially given the nontrivial nature of the histopathological images. Transfer learning implements a more practical solution in a resource-constrained environment, given that it takes a lot less time and computation power given the starting point of a trained neural network. As per the view of Zeng *et al.* (2021), Histopathological images include characteristics like cells, patterns, and textual colours, which are not easily distinguishable by the naked eye. CNN performs feature extraction directly from image input and other pre-training methods at multiple scales. These models can detect delicate components of the tissue, such as edges and textures, which is critical in determining more complex characteristics, including shapes and patterns of tissues necessary for sample classification.

## 2.4 Integration of Multi-modal Data for Improved Diagnostic Precision

Fusing multiple modalities is one of the most significant innovations in detailed diagnostics, especially those healthcare solutions that employ artificial intelligence, machine learning, and deep learning. As Ivanovic *et* *al.* (2023) stated, Integrated data means combining data inputs that arrive in images, text, sensors, and even genealogy for better and more accurate patient diagnostic information. The possibility exists in the complementary relationship of these different types of data; every kind of data yields different information that, if integrated, will improve the diagnostic conclusion and its reliability. For example, intermodal data fusion in medical imaging encompasses the fusion of various images, such as X-ray, CT, MRI, and ultrasound images. CT has specific advantages while MR has others, and so on, among the five kinds of imaging. X-rays, for example, are best suited for producing bone films, CT scan for producing sectional images of organs, MRI for producing high-resolution soft tissue images and ultrasound for determining motion within the body like that of blood or functioning organs. With such data, the clinician is better placed to view the patient from another perspective and may be able to detect deviations or pathological states that may not be recognizable when using any imaging modalities in isolation.

Multi-modal data is also essential in identifying the various machine learning and deep learning models for processing and fusing this data. As per the view of Toms *et al.* (2020), such models can be used to teach programs to identify various patterns of different types of data input. For instance, in cancer screening, MRI images can be integrated with other information, such as genetic markers or history, to identify a tumour's position and information on how fast it may grow or whether it will spread to other body parts. Incorporating genetic information enables one to offer an individualised biochemical diagnosis that considers genetic susceptibilities that may not be revealed by scan. Integrating multi-modal data might be challenging because each modality needs a different type of preprocessing. For instance, image data must be rescaled and standardised in a way; textual data need to be pre-processed and structured; and sensor data may undergo noise filtering.

These preprocessing steps guarantee that the found data is suitable for subsequent ML algorithm-compatible analysis. Also, integrating different types of information may require high-level strategies, which include data fusion, which involves developing algorithms that balance the usefulness of other data sources for diagnosis. As per the opinion of Binte Rashid According to Binte Rashid *et al.* (2024), New insights into using CNNs, RNNs, and transformers in deep learning architectures for processing multimodal data have greatly enhanced image, sequential, and NLP problems. Integrated CNNs with other networks like RNNs and attention-based models allow the system to extract informative features of both image and text input, enhancing the quality of the diagnostic results.

## 2.4 Theoretical Underpinnings

**1. Convolutional Neural Networks (CNNs) for Image Classification**

CNNs are a broad class of deep-learning models that have shown great potential as feature extractors for histopathological image analysis and classification for breast cancer detection. CNNs can adaptively learn spatial pyramids of features from the raw pixel input and can effectively identify and characterise tissue regions and cellular and boundary irregularities of tumours. The different layers of convolutional neural networks (CNNs) are optimized when trained by large amounts of labelled histopathological images and can accurately classify benign and malignant tissues. The model’s layers gradually learn simple features, such as edges, and complex features, such as textures, which are essential for tumour classification (Attallah, 2024). This approach reduces reliance on feature extraction by hand and improves the accuracy of diagnosing cancer at an early stage.

**2. Transfer Learning for Enhanced Model Generalisation**

Transfer learning is a procedure in which a model trained on a broad and largely representative data set is adjusted for a particular task, including breast cancer detection from histopathological images. The use of such an approach is preferable in situations where there are few labelled medical datasets available. Transfer learning is the concept used to allow for better results when making larger, domain-specific models based on the knowledge contained in pre-trained models like VGG or ResNet. Generating a fine set of filters for the model specialised in histopathological images allows it to uncover aspects pertinent to cancer detection. The model becomes more accurate and general depending on the tumour type.

## 2.5 Literature Gap

A vital literature limitation of applying histopathological images of breast cancer to ML models is a lack of a high-quality annotated image data set. This poses a significant challenge for ML models, especially in transferring the models to other populations and clinical settings. More analysis of the use of multi-modal data is needed. All of this could be complemented with other diagnostic data such as genomics, clinical history, or radiology images in the form of histopathological images. The current approaches are mainly confined to image categorization without considering the strengths of multi-dimensional data information (Wang *et al.* 2024). Furthermore, deep learning techniques such as CNNs are dominant. Still, no existing extensive studies compare various categories of ML algorithms in real-world and clinical scenarios, which could help understand the models' robustness, interpretability and clinical relevance.

## 2.6 Chapter Summary

The application of machine learning technology in different stages of breast cancer diagnosis employs histopathological images in developing models that enable sophisticated diagnostic precision and expeditiousness. Convolutional Neural Networks (CNNs) and transfer learning are crucial; CNNs successfully predict intricate tissue imagery classification and transfer learning, boosting performance with limited input data analysis. Nevertheless, obstacles are still present because there are few large, diverse annotated datasets and the incorporation of multi-modal data is also limited. These deficits may be complemented and filled by, for example, increasing and joining various datasets, which can enhance the prognostic usability of these models in clinical practice considerably.

# Chapter 3: Research Methodology

## 3.1 Overview

The aim of this chapter is to describe the methods to investigate the applicability of convolutional neural networks (CNNs) for two-class classification of breast cancer histopathological images. Taken together, the ultimate idea of this chapter is to illustrate and explain in detail the approach that has been taken to conduct this research through document and data gathering, data preparation, model designing and selecting, the method of model training, and the strategy of performance assessment. It comprises a set of consolidated deep learning methods, considerations of the dataset, and accuracy measurements that are applicable to the medical imaging classification problem.

## 3.2 Research Design

This research uses the quantitative experimental research approach focused on the use of predictive modelling in computational pathology. This study aims to study the use of deep learning in classifying breast histopathological images as benign or malignant (Toğaçar *et al.* 2020). The overarching methodology integrates:

* ***Data Collection and Pre-processing***: Sourcing and organizing a curated dataset of breast histopathology images.
* ***Model Development***: Utilizing a transfer learning approach with VGG16, a well-established CNN architecture pre-trained on ImageNet.
* ***Training and Validation***: Implementing stratified partitioning into training and validation sets to ensure robust and unbiased evaluation.
* ***Performance Evaluation***: Employing quantitative metrics (accuracy, loss, precision, recall, F1-score, and confusion matrices).

This structured methodology ensures that the findings are grounded in reproducible, systematic procedures aligned with best practices in machine learning research.

## 3.3 Data Source and Description

The dataset used in the current study includes histopathological images of breast cancer tissue acquired from open access sources (such as the BreakHis dataset or a selected institutional dataset in case the former is not allowed). Every image depict slide of breast tissue which could be developed at any of the four possible power: 40X, 100X, 200X and 400X (Pereira, 2023). For the purpose of binary classification, the dataset is organized into two classes:

* ***Benign***: Histopathology images representing non-malignant tissue.
* ***Malignant***: Histopathology images representing cancerous tissue growth.

**Data Characteristics**:

* ***Image Format***: Typically PNG.
* ***Resolution and Size***: Original sizes may vary; images are resized to a standard resolution (224x224 pixels) to align with the VGG16 input requirements.

**Ethical Considerations**:

All data being used in this study are project and individual identities anonymous, therefore meeting the necessary institutional review board (IRB) criteria (Lehan, 2020). The study does not include contact with patients or any information that would personally identify a patient.

## 3.4 Data Pre-processing

### 3.4.1 Directory Structure and Labelling

To begin with, the histopathological images were arranged into nested directories via their magnification level and class. To reduce the classification problem down into two classes (benign and malignant), the directory structure was reworked. All the images that were identified as benign, whether or not they had been magnified, were moved into an eponymous ‘benign’ folder, while all images that had been identified as malignant were moved into an eponymous ‘malignant’ folder. This approach ensures that the data loading pipeline correctly identifies the dataset as a binary classification issue (Koulali *et al.* 2021).

### 3.4.2 Image Resizing and Normalization

All of the images were reduced to 224 x 224 pixels since it is the expected input dimension for the model being used – VGG16. Pixel intensity levels were pre-processed through scaling between 0 and 1 to address issues of standard deviation and to allow the network to converge.

### 3.4.3 Data Augmentation

To improve generalization and reduce the risk of overfitting, a variety of on-the-fly data augmentation techniques were applied during training, including:

* Horizontal and vertical flips
* Random rotations (up to 20 degrees)
* Zooming (up to 15%)
* Shifting (width and height)
* Shearing transformations

These augmentations simulate variations in tissue appearance and image acquisition conditions, increasing the robustness of the model.

## 3.5 Model Selection: VGG16

### 3.5.1 Rationale for Transfer Learning

Instead of training a CNN from the basics, this research employs transfer learning strategy. VGG16 is the original CNN architecture used for the current research, modified and trained heavily for the ImageNet database. The weights from the ImageNet model offer reasonable out-of-the-box feature learning capacity, particularly in initial-layer convolutions (Bharati and Podder, 2022). This approach is less time-consuming and computationally expensive and sometimes results in better predictive performance on limited data samples.

### 3.5.2 Model Architecture

VGG16 consists of a deep stack of convolutional layers followed by fully connected layers. For this research, the following modifications were made:

* ***Removing Top Layers***: The original top classification layers, designed for the 1,000-class ImageNet task, were removed.
* ***Global Average Pooling Layer***: A global average pooling layer replaced the flattened layer to reduce overfitting and the number of trainable parameters.
* ***Fully Connected Layers***: A new fully connected layer with 512 units and a dropout rate of 0.5 was added to prevent overfitting.
* ***Output Layer***: A single-unit dense layer with a sigmoid activation function produced a binary probability output (Oostwal *et al.* 2021).

At the start of the transfer learning, all the convolutional layers in the VGG16 base were set to be non-trainable to preserve its embedded powerful general-purpose features. After initial training, further fine-tuning of the final few convolutional layers sought to optimize these features to be even more appropriate as pertains the specific task domain.

## 3.6 Training Procedure

### 3.6.1 Data Splits

A stratified split was used to guarantee that the proportion of benign and malignant images was similar in both training and validation sets. In this case, only the ratio of 80:20 for the training and validation datasets, respectively, was applied. The division of data sets maintains that while training a model, its performance is tested exclusively on data sets to comprehend its ability to generalise well.

### 3.6.2 Hyper parameter Settings

* ***Batch Size***: 32 images per batch.
* Learning Rate: Initially set to 1e-4, subsequently reduced to 1e-5 during fine-tuning.
* ***Epochs***: 20 epochs for initial training, followed by an additional 10 epochs after partial unfreezing of the VGG16 layers.
* ***Optimizer***: Adam, chosen for its adaptive learning rate and stable convergence properties.

### 3.6.3 Early Stopping and Regularization

To reduce overfitting, early stopping was carried out using validation loss. Moreover, the learning rate was reduced by using ReduceLROnPlateau when at one point training started to converge to validate that generalization was indeed obtained. The dropout layer and data augmentation also worked as a way of ‘regularization’.

## 3.7 Evaluation Metrics and Techniques

### 3.7.1 Primary Metrics

* ***Accuracy***: The proportion of correctly classified images.
* ***Loss***: Binary cross-entropy loss was monitored to ensure stable training dynamics.

### 3.7.2 Secondary Metrics

To gain a more nuanced understanding of the model’s performance, additional metrics were employed:

* ***Precision and Recall***: Particularly relevant in medical applications where the cost of misclassifications differs for false positives and false negatives.
* ***F1-Score***: The harmonic mean of precision and recall provides a balanced measure when uneven class distributions.
* ***Confusion Matrix***: Allows visualization of true positives, false positives, true negatives, and false negatives, providing insight into model errors (Zhou *et al.* 2020).

## 3.8 Hardware and Software

### 3.8.1 Computing Environment

These experiments were performed on a machine with a GPU as training was done to accelerate the process. The GPU acceleration is important especially considering a computation time is of significant importance when training deep CNNs.

### 3.8.2 Libraries and Frameworks

* ***TensorFlow/Keras***: For building and training deep learning models.
* ***Python (NumPy, Matplotlib, Seaborn)***: For data manipulation, visualization, and statistical analysis.
* ***ImageDataGenerator***: For loading, augmenting, and batching images from directories.

## 3.9 Ethical and Reproducibility Considerations

### 3.9.1 Ethical Use of Medical Data

Although this work only considers aggregated, open-source data, all procedures for data use meet patients’ confidentiality and consent standards as set by data providers. No metadata could identify the patient to somebody who knows him or her well and recognises the photo (Schubbe *et al.* 2020).

### 3.9.2 Reproducibility

For the purpose of reproducing these results, some random seeds were fixed where it was appropriate, and all of the pre-processing steps, model parameters, training regimes, and assessment methods are presented. From the same code snippets and scripts, such as Jupyter notebooks (or the equivalent), it is possible to share the results with other scholars by request, thereby allowing replication or the derivation of additional studies.

## 3.10 Summary

This chapter has described the methodological approach to categorize the breast cancer histopathological images into benign and malignant images. This methodology is transparent and rigorous, as well as replicable because it includes descriptions of the dataset preparation, model architecture, training regimen, and evaluation criteria and computational environment. The subsequent chapters will describe and analyse these experiments and relate the outcomes to prior studies and relevance to practice/practice.

# References

Attallah, O., 2024. Skin-CAD: Explainable deep learning classification of skin cancer from dermoscopic images by feature selection of dual high-level CNNs features and transfer learning. *Computers in Biology and Medicine*, *178*, p.108798.https://doi.org/10.1016/j.compbiomed.2024.108798

Benhammou, Y., Achchab, B., Herrera, F. and Tabik, S., 2020. BreakHis based breast cancer automatic diagnosis using deep learning: Taxonomy, survey and insights. *Neurocomputing*, *375*, pp.9-24.https://doi.org/10.1016/j.neucom.2019.09.044

Bharati, S. and Podder, P., 2022. Machine and deep learning for iot security and privacy: applications, challenges, and future directions. *Security and communication networks*, *2022*(1), p.8951961.https://doi.org/10.1155/2022/8951961

Binte Rashid, M., Rahaman, M.S. and Rivas, P., 2024. Navigating the Multimodal Landscape: A Review on Integration of Text and Image Data in Machine Learning Architectures. *Machine Learning and Knowledge Extraction*, *6*(3), pp.1545-1563.https://doi.org/10.3390/make6030074

Conte, P., Ascierto, P.A., Patelli, G., Danesi, R., Vanzulli, A., Sandomenico, F., Tarsia, P., Cattelan, A., Comes, A., De Laurentiis, M. and Falcone, A., 2022. Drug-induced interstitial lung disease during cancer therapies: expert opinion on diagnosis and treatment. *ESMO open*, *7*(2), p.100404.https://doi.org/10.1016/j.esmoop.2022.100404

Foran, D.J., Durbin, E.B., Chen, W., Sadimin, E., Sharma, A., Banerjee, I., Kurc, T., Li, N., Stroup, A.M., Harris, G. and Gu, A., 2022. An expandable informatics framework for enhancing central cancer registries with digital pathology specimens, computational imaging tools, and advanced mining capabilities. *Journal of Pathology Informatics*, *13*, p.100167.https://doi.org/10.4103/jpi.jpi\_31\_21

Hamida, A.B., Devanne, M., Weber, J., Truntzer, C., Derangère, V., Ghiringhelli, F., Forestier, G. and Wemmert, C., 2021. Deep learning for colon cancer histopathological images analysis. *Computers in Biology and Medicine*, *136*, p.104730.https://doi.org/10.1016/j.compbiomed.2021.104730

Heng, H.P.S., Shu, C., Zheng, W., Lin, K. and Huang, Z., 2021. Advances in real‐time fiber‐optic Raman spectroscopy for early cancer diagnosis: Pushing the frontier into clinical endoscopic applications. *Translational Biophotonics*, *3*(1), p.e202000018.https://doi.org/10.1002/tbio.202000018

Ivanovic, M., Autexier, S., Kokkonidis, M. and Rust, J., 2023. Quality medical data management within an open AI architecture–cancer patients case. *Connection Science*, *35*(1), p.2194581.https://doi.org/10.1080/09540091.2023.2194581

Kora, P., Ooi, C.P., Faust, O., Raghavendra, U., Gudigar, A., Chan, W.Y., Meenakshi, K., Swaraja, K., Plawiak, P. and Acharya, U.R., 2022. Transfer learning techniques for medical image analysis: A review. *Biocybernetics and Biomedical Engineering*, *42*(1), pp.79-107.https://doi.org/10.1016/j.bbe.2021.11.004

Koulali, R., Zaidani, H. and Zaim, M., 2021. Image classification approach using machine learning and an industrial Hadoop based data pipeline. *Big Data Research*, *24*, p.100184.https://doi.org/10.1016/j.bdr.2021.100184

Lehan, T.J., 2020. Continuous improvement of the institutional review board at one completely online university: a transferable framework. *Perspectives: Policy and Practice in Higher Education*, *24*(4), pp.131-135.https://doi.org/10.1080/13603108.2020.1792571

Ma, C., Wang, X., Wu, J., Cheng, X., Xia, L., Xue, F. and Qiu, L., 2020. Real-world big-data studies in laboratory medicine: current status, application, and future considerations. *Clinical Biochemistry*, *84*, pp.21-30.https://doi.org/10.1016/j.clinbiochem.2020.06.014

Oostwal, E., Straat, M. and Biehl, M., 2021. Hidden unit specialization in layered neural networks: ReLU vs. sigmoidal activation. *Physica A: Statistical Mechanics and its Applications*, *564*, p.125517.https://doi.org/10.1016/j.physa.2020.125517

Pereira, M. (2023). BreakHis - Breast Cancer Histopathological Database. *Mendeley Data*, [online] 1. doi:https://doi.org/10.17632/jxwvdwhpc2.1.

Schubbe, D., Scalia, P., Yen, R.W., Saunders, C.H., Cohen, S., Elwyn, G., van den Muijsenbergh, M. and Durand, M.A., 2020. Using pictures to convey health information: A systematic review and meta-analysis of the effects on patient and consumer health behaviors and outcomes. *Patient education and counseling*, *103*(10), pp.1935-1960.https://doi.org/10.1016/j.pec.2020.04.010

Srinidhi, C.L., Ciga, O. and Martel, A.L., 2021. Deep neural network models for computational histopathology: A survey. *Medical image analysis*, *67*, p.101813.https://doi.org/10.1016/j.media.2020.101813

‌Toğaçar, M., Özkurt, K.B., Ergen, B. and Cömert, Z., 2020. BreastNet: A novel convolutional neural network model through histopathological images for the diagnosis of breast cancer. *Physica A: Statistical Mechanics and its Applications*, *545*, p.123592.https://doi.org/10.1016/j.physa.2019.123592

Toms, B.A., Barnes, E.A. and Ebert‐Uphoff, I., 2020. Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems*, *12*(9), p.e2019MS002002.https://doi.org/10.1029/2019MS002002

Wang, J., Jiang, L., Yu, H., Feng, Z., Castaño-Rosa, R. and Cao, S.J., 2024. Computer vision to advance the sensing and control of built environment towards occupant-centric sustainable development: A critical review. *Renewable and Sustainable Energy Reviews*, *192*, p.114165.https://doi.org/10.1016/j.rser.2023.114165

Zeiser, F.A., da Costa, C.A., Roehe, A.V., da Rosa Righi, R. and Marques, N.M.C., 2021. Breast cancer intelligent analysis of histopathological data: A systematic review. *Applied Soft Computing*, *113*, p.107886.https://doi.org/10.1016/j.asoc.2021.107886

Zeng, H., Chen, L., Zhang, M., Luo, Y. and Ma, X., 2021. Integration of histopathological images and multi-dimensional omics analyses predicts molecular features and prognosis in high-grade serous ovarian cancer. *Gynecologic oncology*, *163*(1), pp.171-180.https://doi.org/10.1016/j.ygyno.2021.07.015

Zhou, X., Lu, P., Zheng, Z., Tolliver, D. and Keramati, A., 2020. Accident prediction accuracy assessment for highway-rail grade crossings using random forest algorithm compared with decision tree. *Reliability Engineering & System Safety*, *200*, p.106931.https://doi.org/10.1016/j.ress.2020.106931