BREAST INVASIVE DUCTAL CARCINOMA DETECTION WITH HISTOPATHOLOGICAL IMAGES USING DEEP LEARNING

WINGATES VOON

UNIVERSITI TUNKU ABDUL RAHMAN

BREAST INVASIVE DUCTAL CARCINOMA DETECTION WITH HISTOPATHOLOGICAL IMAGES USING DEEP LEARNING

WINGATES VOON

A project report submitted in partial fulfilment of the requirements for the award of Bachelor of Engineering (Honours) Biomedical Engineering

Lee Kong Chian Faculty of Engineering and Science
Universiti Tunku Abdul Rahman

April 2022

DECLARATION

I hereby declare that this project report is based on my original work except for citations and quotations which have been duly acknowledged. I also declare that it has not been previously and concurrently submitted for any other degree or award at UTAR or other institutions.

Signature :

Name : Wingates Voon

ID No. : 1701174

Date : 4 April 2022

APPROVAL FOR SUBMISSION

I certify that this project report entitled **"BREAST INVASIVE DUCTAL CARCINOMA DETECTION WITH HISTOPATHOLOGICAL IMAGES USING DEEP LEARNING"** was prepared by **WINGATES VOON** has met the required standard for submission in partial fulfilment of the requirements for the award of Bachelor of Engineering (Honours) Biomedical Engineering at Universiti Tunku Abdul Rahman.

Approved by,

Signature : Chai

Supervisor : Ir. Ts. Dr. Hum Yan Chai

Date : 25 April 2022

Signature :

Co-Supervisor :

Date :

ACKNOWLEDGEMENTS

I would like to thank everyone who had contributed to the successful completion of this project. I would like to express my gratitude to my research supervisor, Ir. Ts. Dr. Hum Yan Chai for his invaluable advice, guidance and his enormous patience throughout the development of the research.

In addition, I would also like to express my gratitude to my loving parents and friends who had helped and given me encouragement to dedicate efforts into this project.

ABSTRACT

The staining of haematoxylin and eosin (H&E) in histopathological samples leads to inconsistent colour and intensity variations among digital datasets, thus hindering the performance of deep learning computer-aided diagnostic (CAD) systems. One proposed technique to battle colour invariance among digitalised histopathological images is stain normalisation (SN), which adjusts the source image colour to match the overall colour distribution of other similar images in a dataset. Some studies claimed that SN techniques improved CNNs' performance in histopathological classification tasks, while several contradicted their claims. Therefore, we attempt to justify the importance of SN, specifically Reinhard and Macenko techniques in the invasive ductal carcinoma (IDC) grading application using seven selected CNN models: EfficientNetB0, EfficientNetV2B0-21k, ResNetV1-50, ResNetV2-50, MobileNetV1, and MobileNetV2. Our findings indicated that CNN models trained in the original (non-normalised) dataset outperformed models trained with SN datasets. Among the two SN techniques, the Reinhard average scores topped the Macenko across all evaluation metrics in cross validation (cv) and test results while being more consistent in performance. Hence, we suggest that SN is considered unnecessary to be included in the CNN pre-processing steps to improve CNN performance if effective CNN architectures are employed.

TABLE OF CONTENTS

CHAPTER

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS / ABBREVIATIONS

a	source image in α space
α_1	template image in α space
α_2	processed image in α space
A	vector of the average image
β	source image in *β* space
β_1	template image in *β* space
β_2	processed image in *β* space
B	vector of the dataset image
FN	False negative
G	Number of classes
l	source image in *l* space
l_1	template image in *l* space
l_2	processed image in *l* space
N	total number of classes
N_c	number of classes
N_{sc}	number of samples in each class
S	saturation values of each stain
S_j	other classes output scores
S_p	positive output score
TP	True positive
V	stain vector matrix
w_j	classes weights
A	adenosis
AUC	area under curve
BCG	Breast Cancer Grading Dataset
CAD	computer-aided diagnostic
CNN	convolutional neural network
CV	cross-validation
DC	ductal carcinoma
D_s	source domain

D_t	target domain
EB0	EfficientNetB0
EV2B0	EfficientNetV2B0
EV2B0-21k	EfficientNetV2B0-21k
F	fibroadenoma
FBCG	Four Breast Cancer Grades Dataset
FOV	field-of-view
GANs	generative adversarial networks
GBTC	gradient boosting tree classifiers
H&E	haematoxylin and eosin
HSD	hue-saturation-density
IDC	invasive ductal carcinoma
LC	lobular carcinoma
LN	lymph node
M	Macenko
MC	mucinous carcinoma
MNV1	MobileNetV1
MNV2	MobileNetV2
MRI	magnetic resonance imaging
MT1	Macenko Template 1
MT2	Macenko Template 2
MT3	Macenko Template 3
NGS	Nottingham Grading Scheme
OD	optical density
PC	papillary carcinoma
PT	phullodes tumour
R	Reinhard
ROC	receiver operating characteristic
RT1	Reinhard Template 1
RT2	Reinhard Template 2
RT3	Reinhard Template 3
RV1	ResNetV1-50
RV2	ResNetV2-50
SCD	stain colour descriptor

SN	stain normalisation
SNMF	sparse non-negative matrix factorisation
SPCN	structure-preservinb colour normalisation
SVD	single value decomposition
t-SNE	t-distributed Stochastic neighbour embedding
T1	Template 1
T2	Template 2
T3	Template 3
TA	tubular adenoma
TF2	TensorFlow 2
T_s	source learning task
T_t	target learning task
VLAD	vector of locally aggregated descriptors
WSI	whole slide imaging
WSICS	whole-slide image standardiser

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 General Introduction

Breast cancer, within the vicinity of 2.3 million new cases in 2020, is among the most prevalent diagnosed cancers worldwide (Sung, et al., 2021). Breast cancer is diagnosed when the breast tissue cells proliferate uncontrollably and rapidly, producing a lump in a specific area (American Cancer Society, 2019). The most common type of breast cancer is invasive ductal carcinoma (IDC), accounting for more than 80 % of all occurrences (Sharma, et al., 2010). Early screening and detection are proven to be effective in preventing breast cancer (American Cancer Society, 2019). Therefore, one can screen through (1) mammography, (2) breast magnetic resonance imaging (MRI), and (3) breast ultrasound imaging. In a case where a suspicious lump is discovered, a biopsy is conducted to extract the tissue for further breast cancer analysis — breast cancer grading (IDC grading) (Rakha, et al., 2010).

1.2 Importance of the Study

IDC grading provides an insightful prognosis of a patient's breast cancer condition (Rakha, et al., 2010). IDC grading is one of the primary three prognostic factors affecting breast cancer treatment, with lymph node (LN) condition and tumour size (Shea, Koh and Tan, 2020). Henson, et al. (1991) demonstrated that the prediction accuracy for clinical outcomes improved when both IDC grade and LN condition were utilised together. Frkovic-Grazio and Bracko (2002) showed that the IDC grade predicted tumour behaviour accurately, especially for early small tumours. Schwartz, et al. (2014) showed that high-grade IDC patients who underwent mastectomy suffered higher mortality rates and axillary LN frequency than lower grade patients. Hence, the IDC grade plays a crucial role in determining breast cancer outcomes.

Pathologists evaluate IDC grades based on the Nottingham Grading Scheme (NGS). The NGS criteria include three IDC morphological features:(1) mitotic count (number of tumour cells that are proliferating), (2) nuclear pleomorphism (overall appearance of the tumour cell), and (3) degree of tubule formation (how well the tumour cells reproduce normal glands) (Rakha, et al., 2010). These criteria produce a summation score which is categorised to form grades (Grade 1 to 3), indicating the aggressiveness of the tumour (A lower grade cancer indicates a less aggressive tumour, while a higher grade suggests a more active tumour (Johns Hopkins University, 2021).

1.3 Problem Statement

Manual IDC grading stays as the benchmark for IDC diagnosis currently. However, manual IDC grading is time-consuming and tedious. Manual IDC grading also suffers from high intra- and inter-observation variations among pathologists (He, et al., 2012), with only 75.3 % general agreement (Elmore, et al., 2015). Therefore, computer-aided diagnostic (CAD) systems in histopathological images have developed recently, attempting to overcome the limitations of manual IDC grading (Priego-Torres, et al., 2020).

CAD systems are proven that can reduce interobserver variability while improving efficiency in the automated analysis of histopathological images (Araujo, et al., 2017; Ramadan, 2020). Throughout the years, IDC grading CAD systems have developed from a combination of handcrafted feature extraction methods (Dalle, et al., 2008; Doyle, et al., 2008; Naik, et al., 2008; Basavanhally, et al., 2013; Dimitropoulos, et al., 2017) to deep learning techniques (Wan, et al., 2017; Abdelli, et al., 2020; Li, et al., 2020; Yan, et al., 2020; Senousy, et al., 2021; Zavareh, Safayari and Bolhasani, 2021).

In order to generate digital histopathological images for CAD systems, pathologists are required to prepare the histopathological slides with the following procedures: (1) breast cancer tissue collection, (2) formalin fixation, (3) paraffin section embedment, (4) staining with haematoxylin and eosin (H&E) (Mccann, 2015; McCann, et al., 2015). Finally, the slides are digitised using the Whole Slide Imaging (WSI) methods (Ghaznavi, et al., 2013). The H&E staining technique, a standard staining protocol, highlights the cell nuclei in blue colour (haematoxylin) while other components (cytoplasm and connective tissues) with different pink variations (eosin) (Tellez, et al., 2019).

Nevertheless, H&E images show overlapping regions in the absorption spectrums of multiple stains. As a result, RGB colour transfer can result in undesired colour mixing of the resultant image attributed to the correlation in each colour channel (Ruifrok and Johnston, 2001). Additionally, other factors such as the temperature of the staining solutions, fixation characteristics, imaging device characteristics (Bautista, Hashimoto and Yagi, 2014; Bejnordi, et al., 2016), and slide digitisation conditions (variation in light sources, detectors, or optics) (Veta, et al., 2014) may lead to colour and intensity variation in histopathological images.

Colour variation among digital histopathological images poses a concern in deep learning CAD systems (Zanjani, et al., 2018). The colour difference may negatively affect the model training process (Guo, et al., 2018). Moreover, colour inconsistency in histopathological images may lead to misdiagnosis of malignant cells, as colour facilitates cancer cell detection (Roy, Lal and Kini, 2019). Many advanced deep learning approaches trained with images originating from a source tend to underperform if images from different origins are applied (Goodfellow, et al., 2014; Komura and Ishikawa, 2018; Veta, et al., 2019). Hence, different methods (grayscale and stain normalisations (SN)) are introduced in an attempt to address colour variation in digital histopathological images.

In order to overcome the colour invariance in digital histopathological images, two types of methods have been proposed: (1) converting images into grayscale using different techniques (Hamilton, et al., 1987; Ruiz, et al., 2007; Qureshi, et al., 2008); and (2) standardising images with stain normalisation (SN) techniques (Reinhard, et al., 2001; Macenko, et al., 2009; Bejnordi, et al., 2016; Zanjani, et al., 2018; Lakshmanan, Anand and Jenitha, 2019; Roy, Lal and Kini, 2019; Stanisavljevic, et al., 2019; Lei, et al., 2020). The first solution (grayscale method) provides the average concentration of the tissue constituents but ignores each of their relative amounts and the colour information. Thus, the method is considered infeasible as colour information is proven to be valuable in medical diagnostics (Gupta, et al., 2017; Zanjani, et al., 2018). The second SN approach employs colour modelling to modify the colour of the original image to match the overall colour distribution of other comparable images with or without a template image (Khan, et al., 2014). There are two types of SN methods based on their stain feature extraction techniques: (1) conventional SN methods (Reinhard, et al., 2001; Macenko, et al., 2009; Khan, et al., 2014; Vahadane, et al., 2016) and (2) deep learning-based SN methods (Zanjani, et al., 2018; Shaban, et al., 2019; Lei, et al., 2020; Kang, et al., 2021).

Despite the claimed benefits, some contradictions are found among the literature. Several studies demonstrated that deep learning-based CAD systems, such as the convolutional neural network (CNN), improved when SN techniques were incorporated (Ruifrok and Johnston, 2001; Bejnordi, et al., 2016; Ciompi, et al., 2017; Stanisavljevic, et al., 2019; Munien and Viriri, 2021). For instance, the accuracy of the CNN in classifying H&E-stained colorectal cancer images improved by 20 % when SN was applied (Ciompi, et al., 2017). In addition, the performance of EfficientNets in classifying the ICIAR2018 dataset (Aresta, et al., 2019) improved when SN algorithms were employed during the pre-processing stage (Munien and Viriri, 2021). Furthermore, the CNN accuracy in detecting prostate cancer improved when the SN technique was incorporated into the CNN pre-processing steps (Stanisavljevic, et al., 2019). However, other studies (Gupta, et al., 2017; Telle,z et al., 2019) contradicted the importance of SN. For example, Tellez et al. (2019) claimed that SN is unnecessary to achieve high performance in histopathological image classification; the authors failed to find any substantial performance differences between SN styles but surpassed the greyscale performance. In addition, Gupta, et al. (2017) concluded that employing effective features and classifiers can obviate the need for SN. Therefore, we attempt to justify the importance of SN in the classification of digitalised histopathological images. To the best of our knowledge, we have yet to discover the study of the significance of SN, namely Macenko (Macenko, et al., 2009) and Reinhard (Reinhard, et al., 2001) SN methods in IDC grading application using CNN. Hence, we aim to fill the knowledge gap by providing our findings on the effectiveness of SN methods on automated IDC grading applications with different types of CNN architectures, from simple low-weight CNNs to complex heavy-weight CNNs.

1.4 Aim and Objectives

This study investigates the importance of SN, specifically Macenko and Reinhard techniques in IDC grading with histopathological images using CNN. The objectives are listed below:

i. To validate the role of Macenko and Reinhard techniques in IDC grading application with the publicly available breast cancer grading dataset – Four Breast Cancer Grades (FBCG) Dataset (Abdelli, et al., 2020).

ii. To perform performance analysis on the seven CNN architectures trained with histopathological images with and without SN.

1.5 Scope and Limitation of the Study

This study explored seven types of CNN architectures (EfficientNetB0, EfficientNetV2B0, EfficientNetV2B0-21k, ResNetV1-50, ResNetV2-50, MobileNetV1, and MobileNetV2) to study the importance of SN, namely Reinhard and Macenko SN methods in the automated IDC grading application (see Figure 1.1). Conventional SN methods (Reinhard and Macenko) were included in the CNN pre-processing pipeline, attributable to their wide availability and ease of use. It is acknowledged that conventional SN methods depend on one reference image and may not accurately achieve the style transformation between image datasets (Kang, et al., 2021). Thus, three template images were selected to mitigate this issue. The deep learning-based SN techniques (Zanjani, et al., 2018; Shaban, et al., 2019; Lei, et al., 2020; Kang, et al., 2021) were disregarded due to their robustness, high computational cost, and not being widely available (Kang, et al., 2021). For CNN, the transfer learning technique was adopted to utilise pre-trained TensorFlow 2 (TF2) saved CNNs from TensorFlow Hub for image feature extraction. The saved CNNs were trained on the ImageNet dataset. Next, our proposed method was applied to the Four-Breast-Cancer-Grades (FBCG) dataset. It is important to note that our work was performed without fine-tuning the pre-trained CNN architectures.

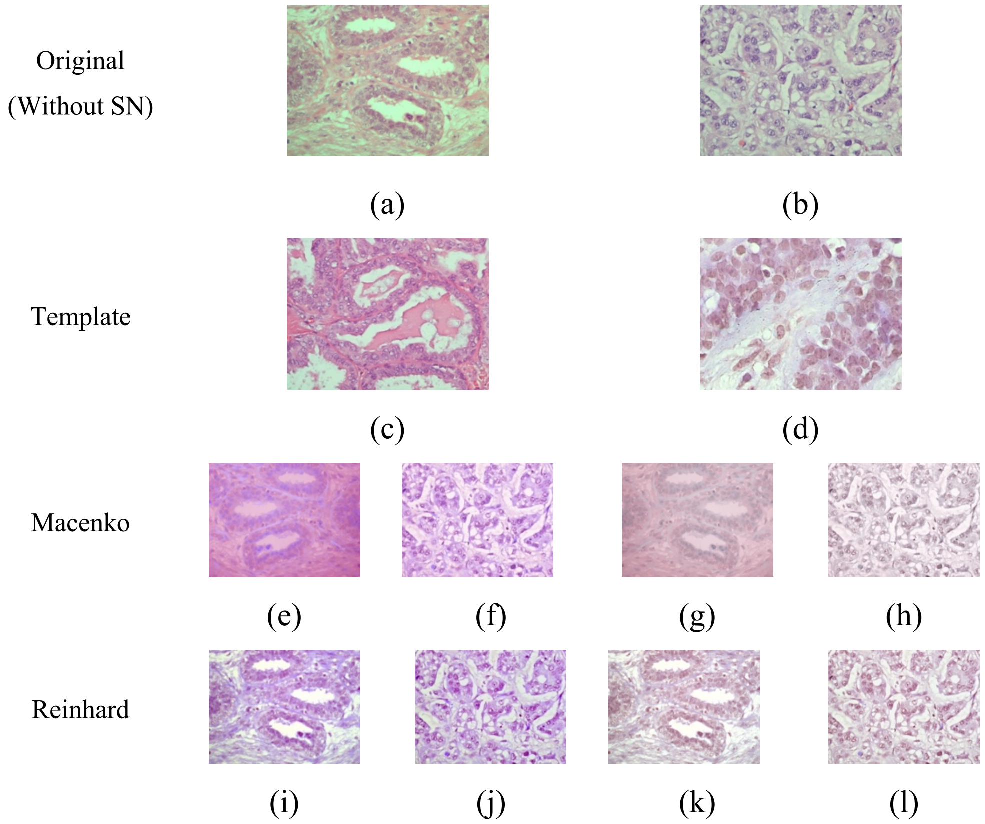


Figure 1.1: Illustration of Images after Macenko and Reinhard Techniques applied to 2 Randomly Selected Images (H&E stained) IDC Histopathological Images (noted that the choice of template image only alters the colour of the original images). (a) Randomly selected H&E stained Image 1; (b) Randomly selected H&E stained Image 2; (c) Randomly selected Template 1 (d) Randomly selected Template 2; (e) Macenko-normalised Image 1 using Template 1; (f) Macenko-normalised Image 2 using Template 1; (g) Macenko-normalised Image 1 using Template 2; (h) Macenko-normalised Image 2 using Template 2; (i) Reinhard-normalised Image 1 using Template 1; (j) Reinhard-normalised Image 2 using Template 1; (k) Reinhard-normalised Image 1 using Template 2; (l) Reinhard-normalised Image 2 using Template 2.

1.6 Contribution of the Study

Two journals related to this topic were successfully submited throughout the project journey. The first journal is titled: "The Breast Cancer Histopathological Images Grading Classification using Convolutional Neural Network Models: A Comparative Study". This journal has received its first revision from Scientific Report, and it is resubmitted currently. The paper source code is available through this link: https://github.com/wingatesv/IDCGradingTask.git. The second journal, titled: "Is Stain Normalisation Important in Breast Invasive Ductal Carcinoma Grading?" is submitted to the Computer Methods and Programs in Biomedicine, awaiting further review. The paper source code is available at this link: https://github.com/wingatesv/StainNormalisationIDCGrading.git. Most importantly, this project won first place in the 2nd IEEE 5 Minutes FYP Competition in 2021. The presentation video is available at this link: https://youtu.be/TnQW-j8xfuw?t=8661. The contributions of this study are summarised as below:

i. This study investigated the importance of SN methods (Macenko and Reinhard) with seven CNN architectures on the IDC grading application.

ii. This study found that selecting the right template image may not necessarily improve the performance of CNN models if ineffective SN is employed.

iii. This study found that SN may be unnecessary to be included in the CNN pre-processing step to improve CNN performance if the effective CNN architecture is used.

1.7 Outline of the Report

This report is structured as follows: In Chapter 2, the development of automated IDC grading systems, the development of SN methods and the SN techniques employed in various breast cancer histopathological images classification tasks are reviewed. Chapter 3 describes the dataset and methodology employed in the study. Chapter 4 presents the findings and results obtained from the experimentations. Finally, Chapter 5 concludes the findings and presents the future works for the study.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This section reviews the development of automated IDC grading applications and SN methods throughout the years. The SN techniques can be divided into two categories: (1) conventional and (2) deep learning-based approaches. Next, this section presents the SN techniques employed in various breast cancer histopathological images classification tasks.

2.2 Automated IDC Grading Systems

This section provides the evolution of automated IDC grading systems from traditional feature extraction techniques to deep learning CNN approaches. For instance, Doyle, et al. (2008) presented an automated quantitative image analysis technique based on spectral clustering and image attributes from the textural and architectural domains. Before conducting spectral clustering, the textural and architectural characteristics from the images were calculated to minimise the feature set's dimensionality. The technique achieved 93.3 % accuracy (low vs high IDC grade) when all architectural factors were incorporated. Basavanhally, et al. (2013) presented a multi-field-of-view (multi-FOV) structure for grading ER+ breast cancers using entire histopathology slides. The authors employed a multi-FOV classification model that can incorporate image characteristics from considerable different sizes of FOVs to predict the breast cancer grade automatically. The approach achieved area under curve (AUC) values of 0.93 (low vs high IDC grades), 0.72 (low vs intermediate IDC grade), and 0.74 (intermediate vs high grades). Dimitropoulos, et al. (2017) presented an automated IDC grading approach by encoding histological images as Grassmann manifold-based Vector of Locally Aggregated Descriptors (VLAD) representations. A new medium-sized breast cancer grading dataset (refer to Breast Cancer Grading (BCG) Dataset onwards) (Zioga, et al., 2017) was created for this study. The study outcome showed that the proposed method achieved an average classification accuracy of 95.8 % when using an overlapping patch size 8x8 strategy.

Nevertheless, these methods are highly feature-based, suffer from high computational power, and lack heuristics for feature extraction (Senousy, et al., 2021). Hence, recent automated IDC grading studies have shifted their techniques towards deep learning (Wan, et al., 2017; Pan, et al., 2020; Yan, et al., 2020; Senousy, et al., 2021).

In automated IDC grading systems, deep learning techniques, particularly CNNs, have become more prevalent (Li, et al., 2020; Yan, et al., 2020; Senousy, et al., 2021). For example, Senousy, et al. (2021) proposed an Entropy-Based Elastic Ensemble of deep CNN models (3E-Net) for breast cancer grading. From the study, the authors utilised multiple CNNs and an ensemble-based uncertainty-measure component to determine the most certain image-wise models for the final breast cancer grading. The proposed models achieved grading accuracy of 96.15 % and 99.50 %, respectively. Despite the success, CNN approaches require much computational power and are more complicated than transfer learning. On the other hand, transfer learning enhances performance by transferring knowledge to a target domain from a source domain (Xu and Dong, 2020). Additionally, transfer learning can reduce time consumption in CNN model training and circumvent the problem of small datasets (Pan and Yang, 2010). The work of Zavareh, Safayari and Bolhasani (2021) employed transfer learning (BCNet) to classify the Databiox (Bolhasani, et al., 2020). The BCNet utilised the VGG16 to extract image features, achieving 88 % validation and 72 % test accuracies in the IDC grading task. Similarly, Abdelli, et al. (2020) devised transfer learning to grade breast cancer using two different CNNs. In the three-breast cancer grade dataset, the MobileNetV1 achieved 93.48 % accuracy, while the ResNetV1-50 achieved 92.39 % accuracy. Additionally, the authors developed a novel dataset strategy (Four-Breast-Cancer-Grades (FBCG) Dataset) by combining the BCG dataset (Zioga, et al., 2017) and BreaKHis (Spanhol, et al., 2016). The study outcome showed that both models performed better on the FBCG dataset than on the original (ResNetV1-50: 97.03 % and MobileNetV1: 94.42 %). Therefore, transfer learning with CNN is proposed to be employed in our study.

2.3 Conventional Stain Normalisation Techniques

Conventional SN methods primarily entail analysing, converting, and matching colour components in histopathological images (Kang, et al., 2021). For instance, the SN method proposed by Reinhard, et al. (2001) (refer to Reinhard onwards) standardises images by matching the source image's statistical colour properties with a template image. Essentially, Reinhard transfers the background colour from the template image to the source images while maintaining other colour intensity information. The authors employed a group of linear transformations, aligning to each CIELAB colour model channel to achieve a unimodal distribution based on the template image. The main drawback of Reinhard assumes that the source and reference images should have identical statistics, which is unlikely for the unique texture property found in each histopathological image (Roy, Lal and Kini, 2019).

Macenko, et al. (2009) devised a stain separation method (refer to Macenko onwards) that automatically locates pixel distribution's fringe in the optical density space. The Macenko mainly comprises three processes: (1) employs the single value decomposition (SVD) to form a plane with the two most significant singular values, (2) projects information to this plane to find the corresponding angles, (3) estimates the stain colour matrix with the maximum and minimum angles for robust stain normalisation. However, the Macenko may generate a poor estimation of the stain vectors in the presence of substantial staining variations (Bejnordi, et al., 2016).

Khan, et al. (2014) conceived a stain normalisation method (refer to Khan onwards) based on the non-linear projection of a source image to a target image using a colour deconvolution representation. First, the stain colour is identified using the Stain Colour Descriptor (SCD). Secondly, a supervised colour classification technique (Relevance Vector Machine) is utilised to determine the position of each present stainThen, these sets of classified pixels estimate the colour formation matrix and stain depth matrix. Likewise, a non-linear spline-based colour normalisation technique is utilised to transform colour to the source image from the target image locally. According to Roy, Lal and Kini (2019), Khan may not entirely maintain the source image histogram structure in the output attributable to the non-linear function.

Vahadane, et al. (2016) presented the structure-preserving colour normalisation (SPCN) method (refer to Vahadane onwards), which deconstructed images into sparse and non-negative stain density maps. Based on the colour of a selected template image, the stain density maps are integrated, changing the colour while maintaining the structure of the source image. The solution space is reduced with the sparse non-negative matrix factorisation (SNMF). Nevertheless, the computation complexity for decreasing the solution space has increased extensively, causing local minima approximation instead of global minima. According to Roy, Lal and Kini (2019), the Vahadane may not preserve the colour variation since only the stain depth matrix was preserved in their study.

Bejnordi, et al. (2016) developed a method known as the whole-slide image colour standardiser (WSICS), which utilises colour and spatial information to categorise the image pixels into distinct stain elements. The chromatic and density distributions of the stain elements in the hue-saturation-density (HSD) colour model are adjusted to fit the template image distribution. WSICS entails six stages: (1) applies HSD conversion; (2) extracts the H&E and background classes automatically from the image while emanating their chromatic and density distributions; (3) alters the 2D chromatic distribution for the dye classes to fit the chromatic distribution of the associated class from a reference image; (4) converts the density distribution for the dye classes to fit the density distribution of the associated class from a reference image; (5) weights the contribution of stains for the pixels and acquiring final chromatic and density conversions; (6) applies inverse HSD conversion.

Roy, Lal and Kini (2019) devised a fuzzy-based modified Reinhard (FMR) colour normalisation approach to control colour coefficients and enhance the contrast of histopathology images. The authors employed fuzzy logic to address the constraints of Reinhard: (1) failed to preserve the background luminance of the source image in the processed image; (2) caused a lower contrast in the processed image when the source image exhibited higher contrast than the template image; (3) caused colour fade due to the transfer of mean colour across all pixels in the image.

Most SN methods (Reinhard, et al., 2001; Macenko, et al., 2009; Khan, et al., 2014; Vahadane, et al., 2016) depend on a template image to approximate stain patterns. Nonetheless, a single template image poses a challenge in including all staining patterns or portraying all input images. As a result, reference image dependency may cause faulty approximation of stain patterns, thus delivering imprecise results (Zhou, et al., 2019; Zheng, et al., 2021).

2.4 Deep Learning-Based Stain Normalisation Techniques

Recently, deep learning-based SN methods have been developed to improve the accuracy of SN without any reference image dependency (Zanjani, et al., 2018; Shaban, et al., 2019; Lei, et al., 2020; Kang, et al., 2021). Zanjani et al. (2018) proposed a new SN method using generative adversarial networks (GANs) to learn the image structures and their association to their colour features. The approach leverages CNNs for non-linear estimation of image distribution in a chromatic space, aligning the colour distribution between source and target image without using statistical properties of the covariance matrix in the chromatic plane. Attributable to the low assumptions of the H&E images attributes, the model is applicable to different types of histopathological images.

Shaban, et al. (2019) presented a deep learning-based unsupervised SN approach (StainGAN) based on CycleGAN (Zhu, et al., 2017) to convert the stain style without needing a reference image. The StainGAN can transfer the H&E stain distribution between different locations without the necessity for paired data from both domains. The StainGAN comprises two generator and discriminator pairs, mapping the images to a domain and then to another domain to maintain image structure. Finally, a similar process is repeated in the reverse direction.

Lei, et al. (2020) presented a deep learning-based SN method (StainCNNs) to speed up the SPCN stain feature evaluation using deep learning. With the TensorFlow framework, the StainCNNs can facilitate the stain feature extraction process while implying a GPU-enabled realisation to increase the learning rate of stain features. Based on the result, the StainCNNs can normalise the whole dataset more efficiently than the SPCN method, which disregards the stain feature distribution in the dataset. However, the StainCNNs assume that the global stain colour distribution of an image is uniform, which may be false in some conditions in histopathological images.

Kang, et al. (2021) designed a new fast and robust deep learning-based SN method (StainNet) to maintain the colour distribution between the source and template images using the distillation learning that can decrease the sophistication of existing deep learning-based SN methods. Without needing a hand-picked reference image, the StainNet can learn the colour distribution from a dataset while modifying the pixel-by-pixel colour value accordingly. The StainNet ensures small size while preventing artefacts in stain conversion. The authors claimed that the StainNet outperformed the StainGAN (Shaban, et al., 2019) by forty times while retaining the source information better and without producing any artefacts.

Compared to conventional SN methods (Reinhard, et al., 2001; Macenko, et al., 2009; Khan, et al., 2014; Bejnordi, et al., 2016; Vahadane, et al., 2016; Roy, Lal and Kini, 2019), deep learning-based SN methods (Zanjani, et al., 2018; Shaban, et al., 2019; Lei, et al., 2020; Kang, et al., 2021) may be outstanding in normalising colour components in histopathological images. However, they tend to suffer from high robustness and computational efficiency. Likewise, the deep learning-based SN methods are usually heavy-weighted, thus, requiring high-computing resources (Zheng, et al., 2021). In addition, It is found that the recently proposed StainNet lacks empirical studies to justify its claim. Therefore, this study proposed to employ the conventional SN methods (Reinhard and Macenko) to investigate the importance of SN in IDC grading applications. The reason this study utilises Macenko, and Reinhard SN methods lie in two folds: (1) publicly available in the python package — StainTools (Byfield, 2020); and (2) well-established among breast cancer histopathology studies (Araujo, et al., 2017; Wan, et al., 2017; Nawaz, et al., 2018; Vesal, et al., 2018; Kassani, et al., 2019; Vo, Nguyen and Lee, 2019; Munien and Viriri, 2021).

2.5 Breast Cancer Histopathological Images Classifications Studies with Stain Normalisation

This section provides studies that utilise SN techniques in breast cancer histopathological images classification tasks. Araujo, et al. (2017) employed CNN (as a feature extractor) with a support vector machine (SVM) classifier to classify the Bioimaging 2015 dataset (Pêgo and Aguiar, 2015). Based on the result, the method achieved 77.8 % accuracy in the 4-class and 80.6 % in the 2-class classification tasks. The authors utilised the Macenko method to normalise the dataset before model training. Wan, et al. (2017) proposed an automated IDC grading method by combining multi-level image features. The authors utilised the Khan SN method before performing nuclei segmentation. The multi-level features extract structural information for accurate cancer morphological classification while cascaded ensembles lower computational costs. The method achieved 92 % accuracy (low vs high grades), 77 % (low vs intermediate grades), 76 % (intermediate vs high grades), and 69 % (low vs intermediate vs high grades).

Likewise, Vo, Nguyen and Lee (2019) developed a hybrid of an ensemble of CNNs and gradient boosting tree classifiers (GBTCs) to classify the Bioimaging 2015 dataset. After pre-processing the dataset with Macenko, three CNNs (Inception-ResNet-v2) were employed to extract visual features and then fed into GBTCs. After merging the GBTCs results with majority voting, the method achieved 96.4 % for the 4-class and 99.5 % for the 2-class classification tasks. The approach taken by Vesal, et al. (2018) proposed to employ transfer learning with Inception-V3 and ResNet50 to classify the ICIAR2018 dataset (Aresta, et al., 2019). The Reinhard SN technique was employed to normalise the dataset prior to model training. The ResNet50 outperformed the Inception-V3 CNN with 94 % accuracy in the classification task. Similarly, Kassani, et al. (2019) conducted classification tasks on the ICIAR2018 dataset (Aresta, et al., 2019) using transfer learning with five different CNNs (Inception-ResNet-V2, Xception, Inception-V3, VGG16, and VGG19). Simultaneously, Macenko and Reinhard SN methods were deployed to study the effect of SN. The result illustrated that the modified Xception CNN trained with Reinhard SN images achieved the highest accuracy (94 %) among other CNNs. From the study, the Reinhard method outperformed the Macenko SN method. Nevertheless, It was discovered that these five studies (Araujo, et al., 2017; Kassani, et al., 2019; Wan, et al., 2017; Vesal, et al., 2018; Vo, Nguyen and Lee, 2019) only employed the SN techniques without observing the performance outcome with the original (non-normalised) dataset.

The following studies have included the performance comparison between CNNs trained with original and SN datasets. Nawaz, et al. (2018) leveraged the power of transfer learning to classify the ICIAR2018 dataset (Aresta, et al., 2019) using a fine-tuned AlexNet. The method scored 81.25 % on validation accuracy and 57 % on test accuracy. Interestingly, the authors discovered an improvement in the Macenko-normalised dataset compared to the original (non-normalised) dataset. Similarly, Munien and Viriri (2021) proposed classifying the ICIAR2018 dataset (Aresta et al., 2019) by leveraging transfer learning with seven EfficientNets. In the study, the authors measured the impact of Reinhard and Macenko SN techniques on the performance of the EfficientNets in the classification task. The study outcome showed that the EfficienNet-B2 achieved the highest result (98.33 %) with the Reinhard SN method among the Macenko SN method (96.67 %) and the original dataset (95.3 %). The study concluded that Macenko and Reinhard SN methods were beneficial in improving the performance of the EfficienNets in classifying the ICIAR2018 dataset. Therefore, it was found that the outcomes of these studies (Nawaz, et al., 2018; Kassani, et al., 2019; Munien and Viriri, 2021) emphasised the advantage of incorporating SN before CNN model training.

However, some contradictions were found in the literature regarding the claimed benefit of SN. Tellez, et al. (2019) investigated the effects of stain normalisation with different histological datasets, including the Camelyon17 Challenge dataset (Bándi, et al., 2019). The authors employed greyscale, Macenko and Bejnordi SN methods with a custom CNN to perform the classification task. The result showed that the SN is unnecessary to achieve high performance in histopathological image classification. Additionally, the authors failed to find any substantial performance differences between SN methods but surpassed the greyscale performance. Gupta, et al. (2017) attempted to answer the question: "Is SN important?" by proposing to classify the BreaKHis (Spanhol, et al., 2016) with grayscale, colour information and Reinhard SN methods. Different texture descriptors and contemporary classifiers were utilised for the classification task. When using colour information (with or without SN), the performance is superior to using grey level information, thus, highlighting the importance of colour in classification. However, with effective features and classifiers, the need for SN may be obviated. It was realised that the importance of SN in CNN applications lacks definite justification. Furthermore, the study on the effect of SN in automated IDC grading applications is even lacking. Therefore, this study aims to fill the knowledge gap by providing our findings on the effectiveness of SN methods (Reinhard and Macenko) on automated IDC grading applications with different types of CNN architectures.

2.6 Summary

Early works of automated IDC grading applications suffered from high computational power, were highly feature dependent and lacked feature extraction heuristics. On the other hand, deep learning, specifically transfer learning, has spurred in the breast cancer histopathology field, thanks to its benefits (time-saving, improved performance and circumventing small datasets issues.) (Pan and Yang, 2010). Therefore, this study proposed to employ transfer learning. The details of these works are summarised in Table 2.1.

Among various SN techniques, deep learning-based SN methods may outperform conventional SN methods in normalising colour components in histopathological images. However, deep learning-based SN methods may be susceptible to high complexity and low computational efficiency. The details of these works are summarised in Table 2.2. Therefore, this study proposed to employ the conventional SN methods (Reinhard and Macenko) to investigate the importance of SN in IDC grading applications, owing to two benefits: (1) publicly available in the python package — StainTools (Byfield, 2020); and (2) well-established among breast cancer histopathology studies (Araujo, et al., 2017; Wan, et al., 2017; Nawaz, et al., 2018; Vesal, et al., 2018; Kassani, et al., 2019; Vo, Nguyen and Lee, 2019; Munien and Viriri, 2021).

After reviewing several studies that include SN techniques in breast cancer histopathological images classification tasks, It was realised that the importance of SN in CNN applications lacks definite justification attributable to the contradiction found in the literature. The details of the literature are summarised in Table 2.3. Moreover, this study has yet to discover the study of the effect of SN, namely Reinhard and Macenko SN methods, in the IDC grading application using CNN. Therefore, This study intends to investigate the effect of SN methods (Reinhard and Macenko) on automated IDC grading applications with different types of CNN architectures.

Table 2.1: Summary of the Development of Automated IDC Grading Applications.

Reference	Method	Dataset	Result
(Doyle, et al., 2008)	Spectral clustering with image textural and architecture features	Custom	93.3 % accuracy with all architecture features
(Basavanhally, et al., 2013)	Multi field-of-view (multi-FOV) classifier	Custom	AUC values: 0.93 (low vs high grades), 0.72 (low vs intermediate grades), 0.74 (intermediate vs high grades)
(Dimitropoulos, et al., 2017)	Grassmann manifold	(Spanhol, et al., 2016; Zioga ,et al., 2017)	95.8 % accuracy (overlapping) patch size 8x8 strategy
(Senousy, et al., 2021)	Deep learning with automatic feature extraction. Entropy-Based Elastic Ensemble of deep convolutional network (CNN) models (3E-Net) for breast cancer grading	(Spanhol, et al., 2016; Zioga, et al., 2017)	3E-Net (Version A): 96.15 % accuracy 3E-Net (Version B): 99.50 %
(Zavareh, Safayari and Bolhasani, 2021)	Transfer learning (feature extraction) with VGG16	(Bolhasani, et al., 2020)	88 % validation accuracy 72 % test accuracy

Table 2.1 (Continued)

(Abdelli, et al., 2020)	Transfer learning (feature extraction) using ResNetV1-50 and MobileNetV1	(Spanhol, et al., 2016; Zioga, et al., 2017)	FBCG dataset: 97.03 % accuracy (ResNet50), 94.42 % accuracy (MobileNet) BCG dataset: 92.39 % accuracy (ResNet50), 93.48 % accuracy (MobileNet)

Table 2.2: Summary of Stain Normalisation Methods over the Years.

Reference	Type	Stain Normalisation	Method
(Reinhard, et al., 2001)	Conventional	Reinhard	Standardises images by aligning the statistical colour distribution (mean and standard deviation) of the source images (CIELAB model channel) with a template image.
(Macenko, et al., 2009)	Conventional	Macenko	Automatically locates pixel distribution's fringe in the optical density space.
(Khan, et al., 2014)	Conventional	Khan	Utilise non-linear mapping of a source image to a target image using a colour deconvolution representation.

Table 2.2 (Continued)

(Vahadane, et al., 2016)	Conventional	Vahadane	Uses SPCN to deconstruct images into stain density maps with SNMF.
(Bejnordi, et al., 2016)	Conventional	Whole-Slide Image Colour Standardiser (WSICS)	Utilises colour and spatial information to categorise the image pixels into distinct stain elements, then adjusting the chromatic and density distributions for the stain components in the HSD colour model to fit the associated distributions from a template image.
(Roy, Lal and Kini, 2019)	Conventional	Fuzzy-Based Modified Reinhard (FMR)	Utilised fuzzy logic to overcome the limitations of Reinhard by controlling colour coefficients and enhancing the contrast of images.
(Zanjani, et al., 2018)	Deep learning	Generative Adversarial Networks (GANs)	Based on GANs to learn the image structures and their relation to their colour attributes for non-linear approximation of image data distribution over chromatic space without relying on statistics of the covariance matrix in a chromatic plane.
(Shaban, et al., 2019)	Deep learning	StainGAN	Based on CycleGAN (Zhu et al., 2017) which comprises two generator and discriminator pairs, maps images to a domain and then to another domain to ensure structure constancy.

Table 2.2 (Continued)

(Lei, et al., 2020)	Deep learning	StainCNNs	Simplify the stain feature extraction in the SPCN method while utilising a GPU-enabled realisation to increase the stain features learning rate.
(Kang, et al., 2021)	Deep learning	StainNet	Uses the distillation learning to retain the colour distribution between the source and target images.

Table 2.3: Review of Breast Cancer Histopathological Studies that employed Stain Normalisation Methods.

Reference	SN Method	Method	Dataset	Result
(Araujo, et al., 2017)	Macenko	CNN with SVM classifiers	(Pêgo and Aguiar, 2015)	4-class: 77.8 % 2-class: 80.6 % Did not study the effect of SN.
(Wan, et al., 2017)	Khan	Deep learning with manual feature extraction. Cascaded ensemble method with multi-level image features combination (pixel, object, semantic)	Custom	92 % (low vs high) 77 % (low vs intermediate) 76 % (intermediate vs high) 69 % (overall) Did not study the effect of SN.

Table 2.3 (Continued)

(Vo, Nguyen and Lee, 2019)	Macenko	Ensemble CNNs (three Inception-ResNet-V2) with GBTCs	(Pêgo and Aguiar, 2015)	4-class: 96.4 % 2-class: 99.5 % Did not study the effect of SN.
(Vesal, et al., 2018)	Reinhard	Transfer learning with Inception-V3 and ResNet-50	(Aresta, et al., 2019)	ResNet50: 97.50 % Inception-V3: 91.25 % Did not observe the outcome of non-normalised dataset.
(Kassani, et al., 2019)	Reinhard and Macenko	Transfer learning with VGG16, VGG19, Inception-Resnet-V2, Xception and Inception-V3	(Aresta, et al., 2019)	Best Model: Xception (94 %) with Reinhard SN technique. Dataset with Reinhard SN generated higher accuracy than the dataset with Macenko SN. Did not investigate the accuracy of non-normalised dataset.

Table 2.3 (Continued)

(Nawaz et al., 2018)	Mecenko	Transfer learning with AlexNet	(Aresta, et al., 2019)	Validation accuracy: 81.25 % Test accuracy: 57 % Performance improvement in the model trained in Macenko-normalised dataset compared to the original (non-normalised) dataset.
(Munien and Viriri, 2021)	Reinhard and Macenko	Transfer learning with seven EfficienNets	(Aresta, et al., 2019)	EfficientNetB2 Reinhard (98.33 %) EfficientNetB2 Macenko (96.67 %) EfficientNets performance better with SN than the original dataset.
(Tellez, et al., 2019)	Greyscale, Bejnordi, Reinhard and Macenko	custom CNN	(Bándi, et al., 2019)	CNN performs better with SN compared to greyscale. CNN performance is not always better with SN.

Table 2.3 (Continued)

(Gupta, et al., 2017)	Greyscale and Reinhard	Seven texture descriptors and Four contemporary classifiers	(Spanhol, et al., 2016)	SN is not needed if given with effective features and classifiers.

CHAPTER 3

METHODOLOGY AND WORK PLAN

3.1 Introduction

This section describes the methodology of this study. In the pre-processing phase, three template images with different conditions: (1) randomly selected, (2) randomly selected from the largest class, and (3) the most average image were selected to mitigate the risk of underperforming SN methods. Subsequently, the FBCG dataset images were stained normalised using the Reinhard and Macenko SN techniques with the selected template image. It is noted that original (non-normalised) images were also included in the experiments to study the impact of SN in automated IDC grading applications. Afterwards, three specific data augmentation techniques (flip, rotation and zoom) were randomly applied to the images to prevent overfitting. For the model training and evaluation phase, each pre-trained CNN architecture was extended with several dropout and dense layers (see Figure 3.8) to perform classification (IDC grading). Finally, each model was evaluated with three performance metrics (balanced accuracy, macro precision, and macro F1-score) to compare the performances in original (non-normalised), Reinhard-normalised and Macenko-normalised datasets. Figure 3.1 depicts the methodology of this study. It was confirmed that all procedures were carried out in accordance with relevant guidelines and regulations.

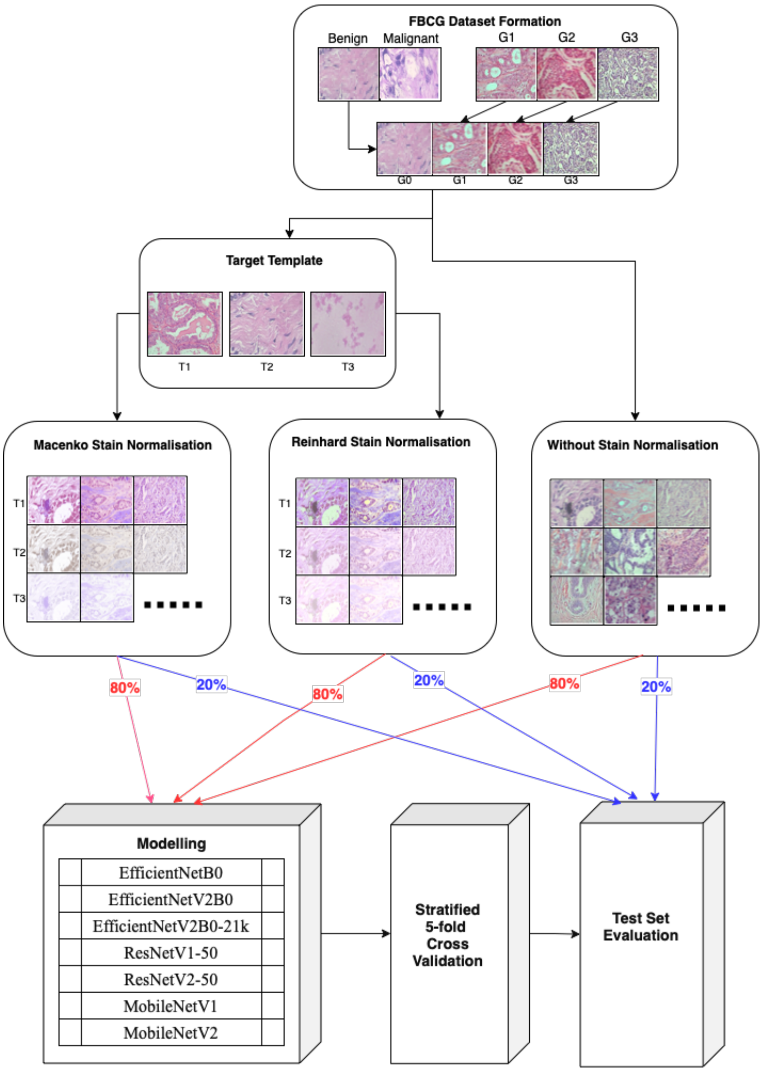


Figure 3.1: The Overall Flow of the Methodology. Initially, a Four-Class Dataset (termed the "Four Breast Cancer Grades (FBCG) dataset") is established using the BreaKHis and BCG datasets.

The dataset images are normalised using Reinhard and Macenko techniques with three different template images. The normalised and the original datasets are fed into seven pre-trained cnn architectures extended with several layers for the IDC grading task. All seven models take the 80 % of the dataset to perform five-fold stratified cross-validation to evaluate the stability of the models. Finally, the seven models are evaluated using the test set (20 % of the dataset).

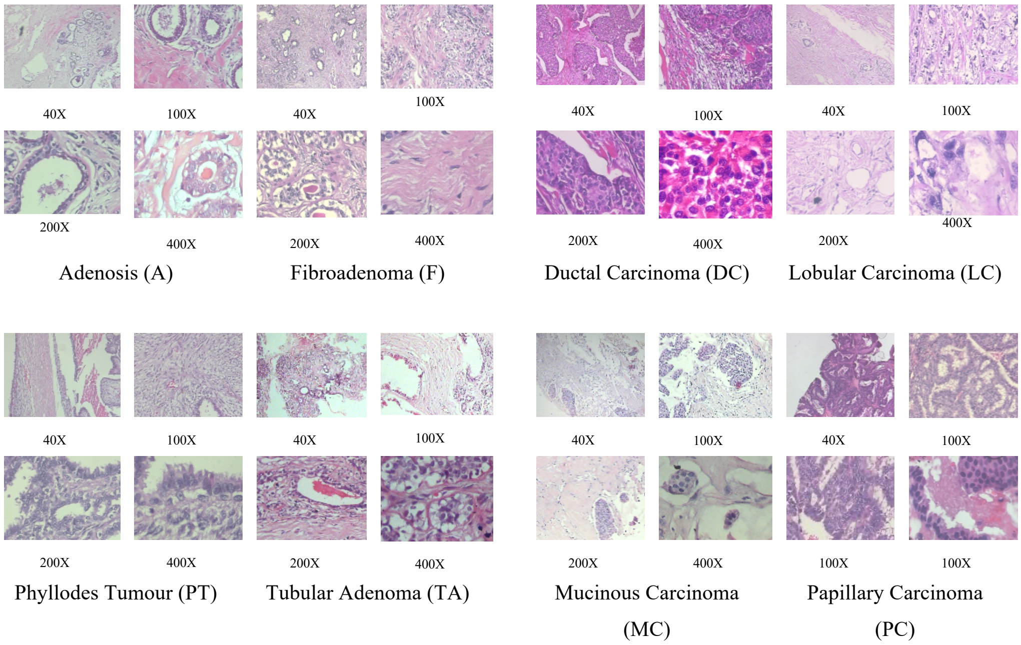
3.2 Dataset

This study employed the FBCG dataset (Abdelli, et al., 2020) to study the importance of SN with seven different CNN architectures in the automated IDC grading application. The FBCG dataset was created by Abdelli, et al. (2020), to address the constraints of small breast cancer datasets. The FBCG dataset entails four classes: (1) Grade 0, (2) Grade 1, (3) Grade 2, and (4) Grade 3. Images of Grade 0 class originate from the BreaKHis dataset (Spanhol, et al., 2016), while Grade 1-3 images are derived from the BCG dataset (Zioga, et al., 2017). The distribution of images in the FBCG dataset is summarised in Table 3.1.

Table 3.1: The Classes (Grade 0-3) Distribution of the FBCG Dataset.

		Grade 0	**Grade 1**	**Grade 3**	**Grade 3**	**Total**
FBCG Dataset	**Train set**	470	86	82	73	711
	Test set	118	21	20	18	177
	Total	588	107	102	91	888

BreaKHis dataset (Spanhol, et al., 2016) was assembled by Spanhol, et al. (2016), containing 7909 breast cancer histopathological images collected from 82 patients. The dataset is mainly split into two categories: benign (2480 images) and malignant (5429 images); benign and malignant breast tumours can be further classified into four distinct types: Adenosis (A), Fibroadenoma (F), Phyllodes Tumour (PT), and Tubular Adenoma (TA) for the benign class; and Ductal Carcinoma (DC), Lobular Carcinoma (LC), Mucinous Carcinoma (MC), and Papillary Carcinoma (PC) (see Figure 3.2). The word "benign" refers to a lesion lacking malignant attributes such as metastasis, significant cellular atypia, mitosis, and disruption of basement membranes. Furthermore, benign lesions are generally non-aggressive, growing slowly, with distinct borders, and localised. However, malignant lesions are often locally invasive and have a proclivity to invade distant sites, resulting in death. The H&E stained breast tissue biopsy slide was firstly captured at four magnification factors (40X, 100X, 200X, and 400X), corresponding to four objective lenses (4X, 10X, 20X, and 40X), then processed into digital RGB format with a resolution of 700 x 460 pixels, generating the digitalised images. Table 3.2 illustrates the distribution of images by class and magnification factor.



(a) Benign (b) Malignant

Figure 3.2: The Images depict H&E Stained Sample Slides with Different Breast Tumour Types in 40X, 100X, 200X and 400X Magnification Factors under Two Major Classes: (a) Benign, and (b) Malignant. This Study regards All Benign Histopathological Images as "Grade 0".

Table 3.2: The BreaKHis Image Distribution by Two Classes (Benign and Malignant) and Four Magnification Factors (40X, 100X, 200X and 400X).

Magnification	Benign	Malignant	Total
40x	625	1,370	1,995
100x	644	1,437	2,081
200x	623	1,390	2,013
400x	588	1,232	1,820
Total	2,480	5,429	7,909

The BCG dataset was published by Zioga, et al. (2017) that contains IDC histological images. The dataset comprises 300 images with three IDC grades: Grade 1 (107 images), grade 2 (102 images), and grade 3 (91 images) that correspond to 21 patients based on their NGS results (see Figure 3.3). The IDC H&E stained histological samples were gathered in the Department of Pathology at Thessaloniki's "Agios Pavlos" General Hospital, Greece, using a Nikon digital camera fitted with a 40X objective lens (equivalent to a magnification of 400X in the BreaKHis dataset), capturing in 1280 x 960 resolution.

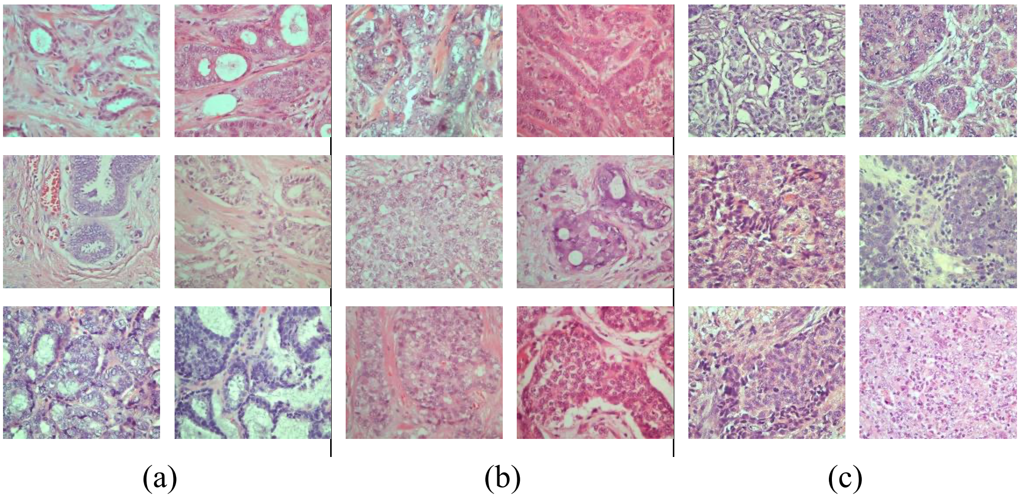


Figure 3.3: These Images are retrieved from the BCG Dataset with Three Classes: (a) Grade 1, (b) Grade 2, and (c) Grade 3.

3.3 Pre-processing

Data pre-processing is essential for histopathological images classification tasks. The CNN is designed to feed on small inputs (Munien and Viriri, 2021). However, the adopted FBCG dataset sizes are much larger (700 x 460 and 1280 x 960) than the seven CNN architectures (see Table 3.3). Thus, the dataset images were shrunk (resized) to ensure that each CNN architecture could receive inputs while preserving the image features. Interestingly, it is worth mentioning that resizing images may preserve global characteristics but ignore local characteristics (Munien and Viriri, 2021). Therefore, the performance of each CNN architecture would highly depend on their ability to recognise and learn global features.

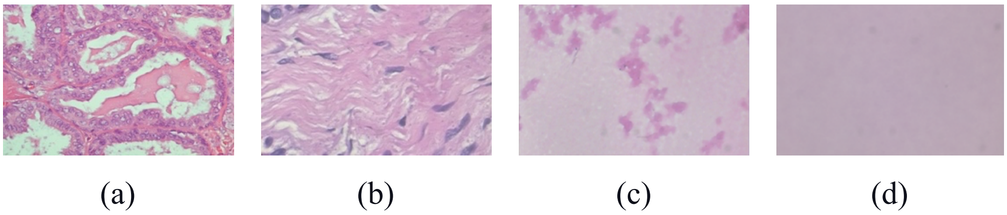
3.3.1 Stain Normalisation

SN techniques (Reinhard, et al., 2001; Macenko, et al., 2009; Bejnordi, et al., 2016; Zanjani, et al., 2018; Lakshmanan, Anand and Jenitha, 2019; Roy, Lal and Kini, 2019; Stanisavljevic, et al., 2019; Lei, et al., 2020) were introduced in an attempt to overcome the limitations of colour inconsistency in deep learning CAD systems, such as (1) reduced model performance (Goodfellow, et al., 2014; Komura and Ishikawa, 2018; Veta, et al., 2019), (2) higher risk of misclassification (Roy, Lal and Kini, 2019) and (3) lower ability of generalisation (Goodfellow, et al., 2014; Komura and Ishikawa, 2018; Veta, et al., 2019). SN generally maps the image colour to modify the source image colour to match the overall colour distribution with or without a template image. Therefore, this study employed the conventional Reinhard and Macenko SN methods to study the effect of SN on the FBCG dataset. Before performing SN, the images were transformed from BGR to RGB colour space to ensure that both SN techniques worked as expected. It is noted that an original (non-normalised) version of the FBCG dataset was included to compare the effect of non-normalised vs normalised datasets (see Figure 1.1).

3.3.1.1 Template Image Selection

This study selected three different template images from three conditions: (1) randomly from the whole dataset, (2) randomly from the largest class (Grade

0), and (3) the image that resembles the most colour average in the dataset based on the cosine similarity method (see Figure 3.4). It is noted that Reinhard and Macenko SN methods rely on one template image and may not accurately accomplish the style conversion between image dataset. In addition, if the chosen template image does not represent the whole dataset, the SN methods may underperform (Kang, et al., 2021). Therefore, selecting three template images allows this study to consider the effect of the template image on SN while mitigating the risk of underperforming SN.



(a) (b) (c) (d)

Figure 3.4: The Selected Template Images for this Study and the Generated Output Image where: (a) Randomly Selected Image from the Dataset, (b) Randomly Selected Image from the Largest Class (Grade 0), (c) the Image that resembles the Most Colour Average based on Cosine Similarity, and (d) the Output of the Average Pixel Intensities of the Dataset.

For selecting the template image that resembles the most colour average in the FBCG dataset, all images in the dataset are converted into arrays of floating points; then, the arrays are summed up to generate the average pixel intensities. Before converting into an 8-bit integer array, the average output value is rounded to the nearest even value. Afterwards, an output image (see Figure 3.4) is generated from the integer array. Finally, the cosine similarity method (see Equation (3.1)) is utilised to compute the similarity between the vector of the output image and each dataset image. The method computes the dot product of the two vectors and divides it by the magnitudes of each vector.

$$\cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{3.1}$$

where

A = vector of the average image

B = vector of the dataset image

3.3.1.2 Reinhard Stain Normalisation

The Reinhard SN technique normalises images by matching the source image's statistical colour properties with a template image. Firstly, both source and template images are read as input data. Secondly, the RGB image is transformed into *lαβ* colour space. The image is converted into independent XYZ space during LMS cone space conversion, followed by converting the XYZ space image to LMS cone space. The skew data in the LMS cone space is eliminated by transforming it into the logarithmic space. Similar to the RGB image, *lαβ* colour space includes its distinct colour (*l, α* and *β* axis denote as achromatic, chromatic blue-yellow, and chromatic green-red channels), which is proportional and compact. Thirdly, *i* = 0 (the number of channels) and *c* = 3 (the number of channels found in the RGB image) are initialised. Then, the condition (if *i* is less than *c*) is applied, followed by the transformation given in the Equation (3.2), (3.3), and (3.4). Finally, the lαβ colour space is converted to an RGB image for display attributable to the incomparable attributes found between *lαβ* colour space and RGB colour space (Reinhard, et al., 2001; Roy, et al., 2018). The flowchart of the Reinhard SM method is illustrated in Figure 3.5.

$$l_2 = mean(l_1) + \big(l - mean(l)\big).* (std(l_1)./std(l)) \tag{3.2}$$

$$\alpha_2 = mean(\alpha_1) + \big(\alpha - mean(\alpha)\big).* (std(\alpha_1)./std(\alpha)) \tag{3.3}$$

$$\beta_2 = mean(\alpha_1) + \big(\beta - mean(\beta)\big).* (std(\beta_1)./std(\beta)) \tag{3.4}$$

where

l_2 = processed image in *l* space

l_1 = template image in l space

l = source image in l space

α_2 = processed image in α space

α_1 = template image in α space

α = source image in α space

β_2 = processed image in β space

β_1 = template image in β space

β = source image in β space

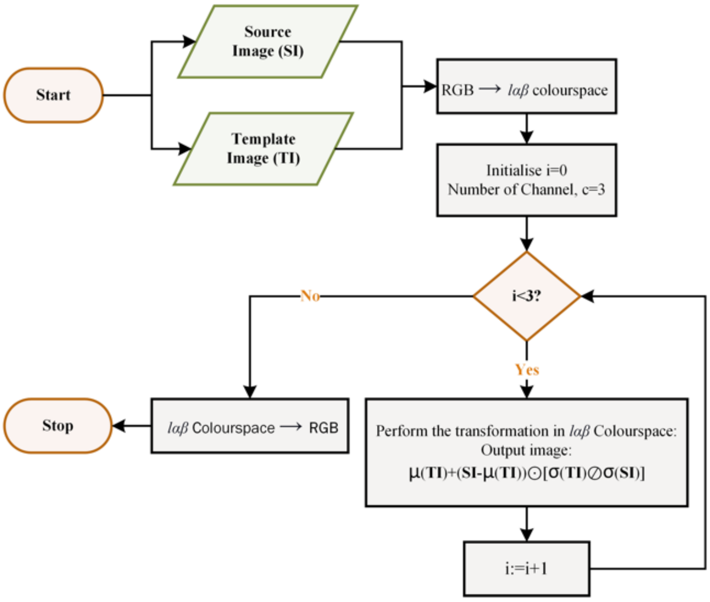


Figure 3.5: The Process of Reinhard SN Technique (Roy, et al., 2018).

3.3.1.3 Macenko Stain Normalisation

The Macenko SN method separates stains automatically by locating the pixel distribution fringe in the optical density space. After reading both source and template images as input, the RGB source image is converted into *lαβ* colour space. Next, the tolerance for the pseudo-minimum αth and pseudo-maximum (100 - α)th percentile is initialised, providing a better yield for further processing at (1) α = 1, (2) optical density (OD) threshold value for translucent pixels β = 0.15, (3) disseminated light intensity I_o = 240, (4) H&E OD matrix and (5) stain concentration. Subsequently, all image colours are transformed into the OD value (see equation (3.5)), offering a linear combination of stains space, thus, resulting in a linear combination of OD values. After applying the condition on the OD threshold value (if β (OD threshold value) < 0.15), the transparent pixels are eliminated so that the OD value is divided into two matrices (see Equation (3.6) and (3.7)) (Macenko, et al., 2009; Roy, et al., 2018).

$$OD = -log_{10}(I) \tag{3.5}$$

$$OD = V * S \tag{3.6}$$

$$S = V' * OD \tag{3.7}$$

where
OD = optical density values
S = saturation values of each stain
V = stain vector matrix

Equations (3.6) and (3.7) locate the stain vector of each image based on the colour (OD value = 0 when the pixel colour is white). In the next step, the singular SVD decomposition value on the OD value is computed. The Geodesic path (Bautista, Hashimoto and Yagi, 2014) is employed to locate the direction where the OD transformed pixel can be projected to locate the final point of the stain vectors. The following process is followed by evaluating the plane formed by vectors (forming a plane with the two vectors associated with the most significant singular value decomposition values of the OD transformed pixel values). Then, All OD values are projected and standardised in the plane. The projected line is plotted, and then the angles in all points to the first singular value decomposition direction are computed, plotting the direction in the plane. The robust extremes fringes in the linear combination of stain vectors are found by calculating the minimum and maximum α^{th} and (100 - α)th percentile. Finally, the H&E stain concentration is determined to the OD values; stain concentration is normalised, then the final image is recreated using a reference mixing matrix (using the H&E matrix with the normalised stain concentration) (Macenko, et al., 2009; Roy, et al., 2018). The flowchart of the Macenko SN method is illustrated in Figure 3.6.

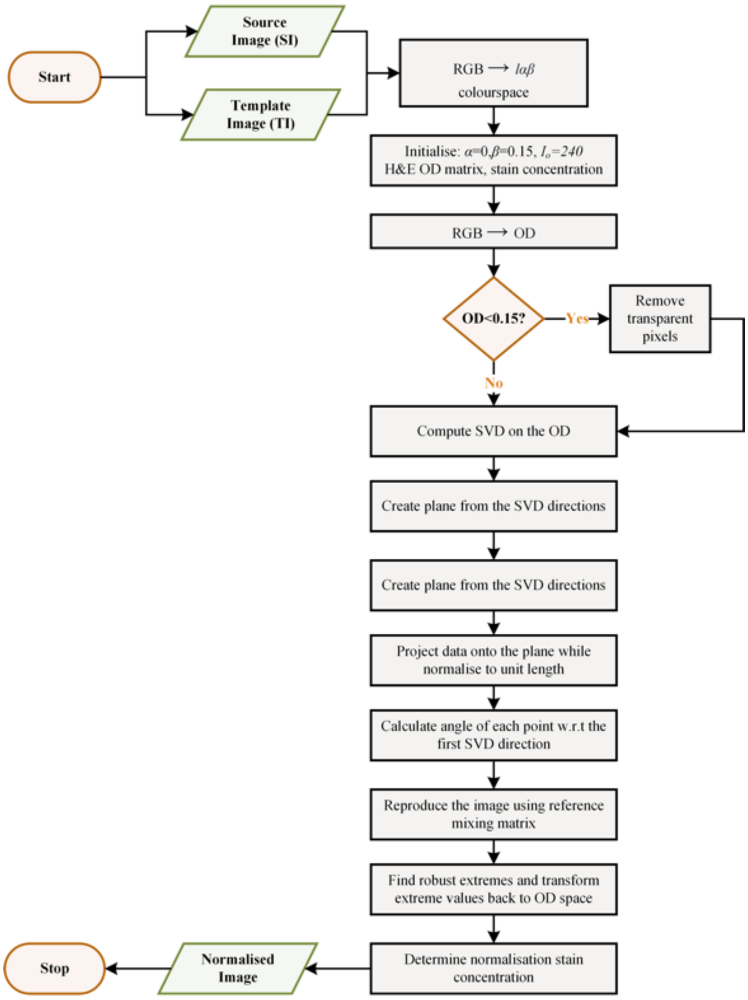

Start
Source Image (SI)
Template Image (TI)
RGB ⟶ lαβ colourspace
Initialise: α=0,β=0.15, I_o=240 H&E OD matrix, stain concentration
RGB ⟶ OD
OD<0.15?
Yes
Remove transparent pixels
No
Compute SVD on the OD
Create plane from the SVD directions
Create plane from the SVD directions
Project data onto the plane while normalise to unit length
Calculate angle of each point w.r.t the first SVD direction
Reproduce the image using reference mixing matrix
Find robust extremes and transform extreme values back to OD space
Determine normalisation stain concentration
Normalised Image
Stop

Figure 3.6: The Process of Macenko SN Technique (Roy, et al., 2018).

3.3.2 Data Augmentation

Data augmentation is essential in mitigating overfitting risk during model training, especially if the employed dataset is delimited (Shorten and Khoshgoftaar, 2019). Hence, the training data was infused with artificial diversity via random but realistic transformations to address the overfitting issue. Specifically, TensorFlow Keras pre-processing layers were employed to augment the training data (the layers are disabled automatically during model validation and testing). Three data augmentation techniques were utilised: (1) random horizontal and vertical flips, (2) random rotation and (3) random zoom (see Table 3.4). Random flipping and rotation were employed because the pathologists' ability to examine histopathological images is not affected by rotation angles. Hence, it was assumed that different rotation angles would not affect CNN's learning ability. Additionally, random zooming was employed to imitate the magnification factor found in histopathological images to improve the CNN's generalisation ability. Figure 3.7 shows the samples of random data augmentation.

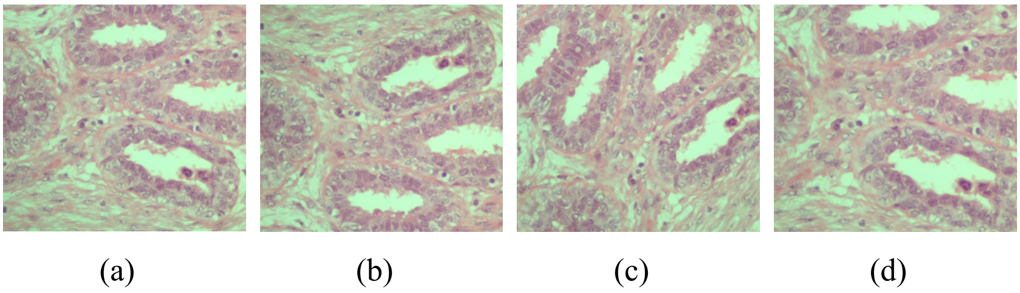


Figure 3.7: These Images are the Sample of Random Data Augmentation where: (a) Original Image, (b) Flipped Image, (c) Rotated Image, and (d) Zoomed Image.

3.3.3 Data Balancing

The employed FBCG dataset for this study suffers from a data imbalance issue (see Table 3.1). If the problem is unresolved, the model tends to be more biased toward predicting the majority class (Grade 0). Therefore, this study implemented the class weighting technique from the Scikit-Learn Python library to resolve this concern. This technique grants the minority class a higher weight in the model cost function to impose a more significant penalty on the minority class. As a result, the model can converge to minimise errors for the minority class (Analytics Vidhya, 2020). The following equation was employed to determine the weight of each class:

$$W = \frac{N}{N_c \times N_{sc}} \tag{3.8}$$

where

$W =$ class weight

$N =$ total number of samples

$N_c =$ number of classes

$N_{sc} =$ number of samples in each class

3.4 Transfer Learning

This study employed transfer learning attributable to three motivations: (1) improved CNN performance, (2) time savings, and (3) being able to mitigate small datasets issue. In general, transfer learning entails four phases: (1) source domain (D_s), (2) target domain (D_t), (3) source learning task (T_s), and (4) target learning task (Tt). Transfer learning attempts to enhance the target predictive function $D_t(.)$in D_t with the knowledge in D_s and T_s, where $D_s \neq D_t$ *or* $T_s \neq T_t$ (Pan and Yang, 2010). Commonly, a CNN is constructed so that its first several layers learn more generic features (edges and generic shapes), while the last several layers recognise specific features related to the problem. Thus, the transfer learning technique utilises the general features learned in the first few layers of the source dataset and then relearns the specific features in the target dataset in the several final layers (Xu and Dong, 2020).

Transfer learning entails two distinct methods for customising a pre-trained model: (1) feature extraction and (2) fine-tuning. Feature extraction leverages a previous network's representations to extract critical features from a new dataset by superimposing new classifier layers on top of the pre-trained model, repurposing the previously learned feature representations on the new dataset. Contrarily, fine-tuning unfreezes several top layers of the pre-trained model while training both newly added classifier layers and the unfrozen layers of the pre-trained model. Fine-tuning the pre-trained model's specific feature representations (high-order features) may make the representations more applicable for a particular task (TensorFlow, 2021). Although fine-tuning may improve model performance, this technique may induce overfitting. Hence, this study employed feature extraction using the seven pre-trained CNN architecture. Each pre-trained CNN architecture was utilised in the form of an image feature vector (a dense 1D tensor describing the whole image), reposited in the TensorFlowHub (TensorFlow, 2022). To apply the feature vector to our work, the "hub.KerasLayer" was employed to integrate the feature vector into our framework. This layer produces a batch of feature vectors whose size is proportional to the input size. The seven CNN architectures used in this study is summarised in Table 3.3.

Table 3.3: Summary of the Seven Pre-Trained CNN Architecture employed in this Study in terms of their Contribution, Trained Dataset, Flops, Parameters and Input Shape.

Architecture	Contribution	Dataset	FLOPs (B)	Parameters (M)	Input Shape
EfficientNetB0 (Tan and Le, 2019)	Compound scaling	ImageNet - ILSVRC-2012-CLS)	0.39	5.3	224 x 224
EfficientNetV2B0 (Tan and Le, 2021)	Progressive learning	ImageNet - ILSVRC-2012-CLS	0.72	7.1	224 x 224
EfficientNetV2B0-21k (Tan and Le, 2021)	Progressive learning	ImageNet-21k	0.72	7.1	224 x 224
ResNetV1-50 (He, et al., 2015)	Residual learning	ImageNet - ILSVRC-2012-CLS	4.1	25.6	224 x 224
ResNetV2-50 (He, et al., 2016)	Identity mapping	ImageNet - ILSVRC-2012-CLS	4.1	25.6	224 x 224
MobileNetV1 (Howard, et al., 2017)	Depth-wise separable convolutions	ImageNet - ILSVRC-2012-CLS	0.6	4.2	224 x 224
MobileNetV2 (Sandler, et al., 2018)	Inverted residuals and linear bottlenecks	ImageNet - ILSVRC-2012-CLS	0.3	3.4	224 x 224

3.5 Experiment Details

This study conducted the experiments in the Google Collaboratory, with the specifications: (1) 2.30GHz Intel (R) Xeon (R) CPU, (2) 12GB RAM, (3) up to 358GB disc space, and (4) 12GB/16GB Nvidia K80/T4 GPU. The FBCG dataset was divided into train-test sets with a spilt of 80 %-20 % without overlapping. the test set images were selected through the stratification process by extracting the selected first portion of images in the dataset. The distribution of images in the FBCG dataset is summarised in Table 3.1. It is noted that the images in the dataset were stained normalised (Reinhard and Macenko techniques) using the StainTools (Byfield, 2020) python package. Next, the train set was further divided into five folds to accomplish the stratified five-fold cross-validation (CV). The stratified CV ensures that each training set fold acquires identical observations with a given label while ensuring that each CNN model is appropriately trained.

After creating the CV folds, the Keras ImageDataGenerator (from TensorFlow Keras preprocessing.image) was employed along with its designated function (flow_from_dataframe) to generate batches of processed (rescaled, resized, and shuffled) input data. It is noted that the seed value was remained the same during experimentations to ensure training and validation data remained identical. The implementation details are summarised in Table 3.4. To implement each pre-trained CNN architecture, the corresponded feature vector publicly available from the TensorFlowHub was deployed. The feature vector was integrated with several other layers (input layer, data augmentation layer, drop out layers and dense layers) to form a CNN model. The final structure of the model entails seven layers:

i. An input layer.
ii. A data augmentation layer.
iii. The feature vector.
iv. A dropout layer(rate = 0.5) .
v. A dense layer (256 neurons and ReLU activation function).
vi. A dropout layer (rate = 0.4).
vii. A dense layer (4 neurons and SoftMax activation function).

The seven adopted pre-trained CNN architectures were standardised with the same framework and hyperparameters to ensure a fair comparison. The architecture of the model framework is depicted in Figure 3.8. First, the input layer assigns a specific shape to the input data (image resolution). Second, the data augmentation layer augments (randomly flips, rotates, and zooms) the input data in model training. After data augmentation, the input data is fed into the feature vector to extract general features. Then, the output data flows through the first dropout layer (rate = 0.5), the fully connected layer (256 neurons), the second dropout layer (rate = 0.4), and the output fully connected layer (4 neurons). Finally, the SoftMax function converts the model output to a vector of probabilities for each class's input data (A dropout rate describes the rate at which input units are assigned to 0 in a dropout layer. If the input units are not assigned to 0, they are scaled up by 1/(1-rate) to maintain the same sum of all inputs (Keras, 2021)).

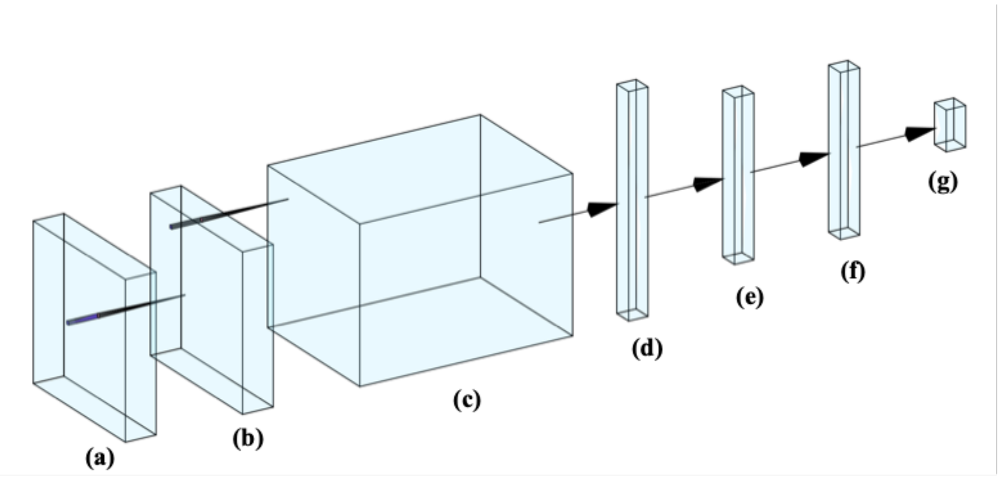


Figure 3.8: The Standardised Model Framework for this Study where: (a) Input Layer, (b) Augmentation Layer, (c) Feature Vector (Pre-Trained CNN), (d) Dropout Layer (Rate = 0.5), (e) Dense Layer (256 Neurons), (f) Dropout Layer (Rate = 0.4), and (g) Dense Layer (4 Neurons).

Each model was compiled with the Adam Optimiser (learning rate = 0.001). Deciding on a suitable learning rate is essential for model training since it influences the time required for the model to converge to local minima. A rapid learning rate may render the model deviate from its local minima. Contrarily, a slow learning rate may impede model training, resulting in increased computational cost (Zulkifli, 2018). Therefore, after several empirical tests, this study selected the 0.001 learning rate as the optimal value. Likewise, the weighted categorical cross-entropy loss function was implemented for the classification task that utilises the weight class technique and the metrics parameter "accuracy." Finally, each fold was trained for 100 epochs. The details of the model's compilation are summarised in Table 3.4. The weighted categorical cross-entropy loss function is defined as:

$$WCE = -w_j * log\left(\frac{e^{s_p}}{\sum_j^c e^{s_j}}\right) \tag{3.9}$$

where

$WCE =$ weighted categorical cross-entropy

$S_p =$ positive output score

$S_j =$ other classes output scores

$w_j =$ classes weights

Table 3.4: Summary of the Data Pre-Processing, Data Augmentation, And Model Compilation Details for this Study.

	Parameters	**Values**
Pre-processing (flow_from_dataframe)	target_size	N x N (see Table 3.3)
	batch_size	16
	shuffle	True
	seed	123
	class_mode	categorical
Data Augmentation	RandomFlip	horizontal_and_vertical
	RandomRotation	0.2
	RandomZoom	0.2

Table 3.4 (Continued)

Model Compilation	Optimiser	Adam Optimiser
	Learning rate	0.001
	Loss function	Weighted Categorical Cross Entropy
	Metrics	Accuracy
	Epochs	100

3.6 Performance Evaluation Metrics

For this study, three evaluation metrics were employed to assess the performance of the seven CNN architectures due to data imbalance: (1) balanced accuracy, (2) macro precision, and (3) macro F1-score. Adapted from the Scikit Library, the balance accuracy score calculates the average of recall acquired in each class. Next, the macro-average technique calculates each class metric independently and averages the results, thus, ensuring that all classes are treated equally. It is noted that the standard evaluation metrics (accuracy, precision, recall and F1-score) were disregareded since the adopted FBCG dataset is imbalanced. Finally, the model's ability to differentiate between classes was quantified with the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) (Bex, 2021). The following mathematical expressions define the evaluation metrics:

$$Balanced\ Accuracy = \frac{1}{|G|}\sum_{i=1}^{|G|}\frac{TP_i}{TP_i + FN_i} \tag{3.10}$$

$$Precision_{\ macro} = \frac{1}{|G|}\sum_{i=1}^{|G|}\frac{TP_i}{TP_i + FP_i} \tag{3.11}$$

$$F1_{macro} = 2\frac{Precision_{macro} \times Balanced\ Accuracy}{Precision_{macro} + Balanced\ Accuracy} \tag{3.12}$$

where

TP = True positive

FN = False negative

G = Number of Classes

3.7 Summary

This study employed the FBCG dataset, which entails four classes (Grade 0, Grade 1, Grade 2, and Grade 3) for the automated IDC grading application (see Table 3.1). All the images in the dataset were normalised with Reinhard and Macenko SN techniques (see Figure 1.1) using StainTools (Byfield, 2020). The Reinhard SN technique normalises images by matching the source image's statistical colour properties with a template image (see Figure 3.5). Likewise, the Macenko SN method separates stains automatically by locating the pixel distribution fringe in the optical density space (see Figure 3.6). Both SN techniques require a template image; thus, three template images from different conditions were selected to reduce the risk of underperforming SN techniques. The original (non-normalised) dataset was included in the experiment to compare the effect of SN. Subsequently, three data augmentation methods (flip, rotation and zoom) were applied to the images using the Keras pre-processing layers before model training (see Figure 3.7).

The FBCG dataset was divided into an 80 %-20 % train-test split; the train set was split into the stratified five-fold CV. The images were processed into batches with the Keras ImageDataGenerator before feeding into each pre-trained CNN architecture (see Table 3.3) extended with several dropout and dense layers (see Figure 3.8) to perform classification. The implementation details are summarised in Table 3.4. Finally, each model was evaluated with three performance metrics (balanced accuracy, macro precision, and macro F1-score) to compare the results.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction

This section provides the performance evaluations (balanced accuracy, macro precision and macro F1 scores) of the seven pre-trained CNN architectures trained in the original, Reinhard-normalised, and Macenko-normalised FBCG datasets with three different template images. The illustrations of accuracy and loss curves, confusion matrices, ROC curves and the visual comparison of the t-distributed Stochastic neighbour embedding (t-SNE) of the best and worst CNN models are provided for further analysis. The limitations and challenges of the study are elaborated to describe the scope and difficulties faced in the experimentations.

4.2 Analysis of Results

Seven state-of-the-art pre-trained CNN architectures: 1) EfficientNetB0, EfficientNetV2B0, EfficientNetV2B0-21k, ResNetV1-50, ResNetV2-50, MobileNetV1, and MobileNetV2 were employed to classify the original, Reinhard-normalised and Macenko-normalised FBCG datasets with three different template images. Based on Table 4.1, Table 4.3 and Table 4.5, SN methods referred to Template 1 (T1) image outperformed other template images across all average metrics (0.8917 ± 0.0277, 0.8814 ± 0.0307, 0.8829 ± 0.0297) in cv results. On the other hand, the Template 2 (T2) image achieved the highest average score across all metrics (0.8835 ± 0.0239, 0.8753 ± 0.0277, 0.8762 ± 0.0252) among other template images for test results. Among the CNN models trained on SN T1 datasets, the ResNetV1-50 achieved the highest score across all metrics in the Macenko T1 stained cv results. Likewise, the EfficientNetB0 ranked as the top achiever in balanced accuracy (0.9239), whereas the ResNetV2-50 outperformed all CNN models in macro precision (0.9063) and macro F1-score (0.9060) in Macenko T1 stained test results. In SN T2 stained cv results, the MobileNetV1 scored the top across all metrics (0.9208 ± 0.0292, 0.9253 ± 0.0425, 0.9212 ± 0.0343) when stained using the Macenko technique.

However, the EfficientNetV2B0-21k outperformed other CNN models in balanced accuracy (0.9294) and macro F1-score (0.9117), while the MobileNetV1 scored the highest macro precision (0.9041) in Macenko T2 stained test results. For the worst-performing Template 3 (T3), the EfficientNetV2B0-21k appeared to be the top scorer across all metrics (0.8840 ± 0.0424, 0.8805 ± 0.0281, 0.874 ± 0.0387) in Reinhard T3 stained cv results whereas the ResNetV2-50 achieved the top scores across all metrics (0.9030, 0.9012, 0.8992) in Reinhard T3 stained test results.

Although most top-performing CNN models were being trained on Macenko-normalised datasets, the average scores for the Reinhard technique across all metrics in cv and test results still outperformed the Macenko technique. Hence, this finding suggests that the Reinhard technique performs more consistently than the Macenko technique. Among the three template images, this study hypothesized that T3 would outperform T1 and T2 images since T3 resembles the most colour average image found in the dataset and, thus, able to represent the whole dataset. Nevertheless, both T1 and T2 images achieved better performances than T3, proving our hypothesis otherwise. Therefore, this study suggests that selecting the right template image may not necessarily improve the performance of CNN models if ineffective SN is employed.

Table 4.2, Table 4.4, and Table 4.6 illustrate the balanced accuracy, macro precision, and macro F1-score of the seven CNN models acquired from the five-fold stratified cv and test sets trained in the original, Reinhard and Macenko FBCG datasets. The experimental results show that the highest overall performance of the seven CNN models was achieved on the original FBCG dataset instead of SN. The average score ranked the highest in all performance metrics when the original dataset was used for model training, followed by Reinhard and the Macenko technique. Overall, the EfficientNetV2B0-21k achieved the top score across all metrics (0.9666 ± 0.0185, 0.9646 ± 0.0174, 0.9642 ± 0.0184) in the cv results and the balanced accuracy (0.9524) in test results with the original dataset. Likewise, the MobileNetV1 scored the highest across all metrics (0.9545, 0.9524, 0.9487) in test results with the original dataset. When the FBCG dataset was stained using the Reinhard technique, the EfficientNetV2B0-21k achieved the highest score across all metrics (0.9058 ± 0.0196, 0.8970 ± 0.0170, 0.8971 ± 0.0206) in cv results.

However, in the test results, the ResNetV1-50 emerged as the top performer across all metrics (0.9027±0.0077, 0.8945±0.0175, 0.8944 ± 0.0161). Similarly, the ResNetV1-50 achieved the highest score across all metrics (0.8917 ± 0.0423, 0.8817 ± 0.0490, 0.8822 ± 0.0484) in cv results, macro precision (0.8691 ± 0.0409) and macro F1-score (0.8706 ± 0.0301) in test results when the model was trained on Macenko-normalised FBCG dataset. For the balanced accuracy of test results, the EfficientNetB0 ranked as the top performer (0.8795 ± 0.0461). The Reinhard technique outperformed the Macenko technique in the IDC grading application, aligning with the findings published by this study (Munien and Viriri, 2021).

Nevertheless, CNN models trained in the original FBCG dataset outperformed both SN techniques (see Figure 4.1). From the bar charts in Figure 4.2 and Figure 4.3, a significant performance disparity can be observed in the IDC grading classification performance using images without SN compared to those CNN models using Reinhard and Macenko techniques. Furthermore, it is found that the results generated from CNN models trained without SN show the highest stability compared to those trained with SN (see Figure 4.4). Our result contested the general presumption that SN is essential to accomplish top performance in the histopathological classification tasks, similar to the results published by Tellez, et al. (2019). Therefore, this study considers that SN may be unnecessary to be included in the CNN pre-processing step to improve CNN performance if the effective CNN architecture is used.

The graphs in Figure 4.5 depict the accuracy and loss curves of the best and worst-performing CNN models in the original and SN FBCG test sets. The curves describe the process of the training and validation accuracy and loss (per epoch) in the CNN models training. No sign of model overfitting is observed since the validation accuracy curves are higher than the training accuracy curves, whereas the validation loss curves are lower than the training loss curves. The validation loss curves are significantly lower than the training curves are most likely the result of implementing dropout layers in the CNN model frameworks. Another reason for the gap is that the CNN models find the validation set easier to predict than the training set, attributed to its unrepresentativeness of the whole dataset (Munien and Viriri, 2021).

Figure 4.7 illustrates the normalised confusion matrices of the best and worst-performing CNN models in the original and SN FBCG test sets. The matrices demonstrate that all the chosen CNN models achieved the highest performance in classifying Grade 0 but the lowest in identifying Grade 1 except for the Macenko T3 EfficientNetV2-B0.

Figure 4.6 illustrates the ROC curves of the best and worst-performing CNN models in the original and SN FBCG test sets. The ROC curves are generated by computing and generating the true positive rate versus the false-positive rate for a binary classifier over a range of threshold values. The area under the curve (AUC) values of the figures show that all chosen CNN models demonstrate the highest performance in identifying Grade 0 and the lowest performance in identifying Grade 1 except for the Macenko T3 EfficientNetV2-B0. The original EfficientNetV2-B0-21k (top left) achieved the highest overall AUC scores, whereas the Macenko T3 EfficientNetV2-B0 scored the lowest overall AUC score.

Figure 4.8 shows the visual comparison of the distribution of learned features using t-distributed Stochastic neighbour embedding (t-SNE) of the best and worst-performing CNN models in the original and SN FBCG test sets. The t-SNE is a statistical approach for visualising high-dimensional information by assigning a position to each point on a two dimensional map. The figure shows that the CNN models can separate different grading explicitly into distinct groupings based on characteristics extracted from the models. The t-SNE revealed that the SN has indistinct effect on the separation of the classes compared to the original FBCG dataset, thus supporting that SN is unnecessary in the CNN pre-processing step to improve CNN performance.

Table 4.1: Balanced Accuracy Results of Seven CNN Models trained in Reinhard (denoted as R) and Macenko (denoted as M) FBCG Datasets with Three Different Templates (denoted as T1, T2 and T3).

SN	**Models**	**Balanced Accuracy**					
		T1		**T2**		**T3**	
		CV (mean±std)	**Test**	**CV (mean±std)**	**Test**	**CV (mean±std)**	**Test**
R	**EfficientNetB0**	0.8777±0.0310	0.8975	0.8702±0.0516	0.8877	0.8202±0.0497	0.8257
	EfficientNetV2B0	0.8845±0.0307	0.8772	0.8524±0.0503	0.8871	0.7764±0.0509	0.7828
	EfficientNetV2B0-21k	**0.9218±0.0393**	0.8844	**0.9116±0.0377**	0.8469	**0.8840±0.0424**	0.7746
	ResNetV1-50	0.8990±0.0197	**0.8983**	0.8921±0.0415	**0.9115**	0.8542±0.0622	0.8990
	ResNetV2-50	0.8975±0.0259	0.8955	0.8814±0.0369	0.8865	0.8470±0.0672	**0.9030**
	MobileNetV1	0.9043±0.0299	0.8345	0.8740±0.0737	0.8661	0.8513±0.0691	0.8333
	MobileNetV2	0.8812±0.0399	0.8509	0.8898±0.0514	0.8897	0.8673±0.0364	0.8641

Table 4.1 (Continued)

	Average (mean±std)	**0.8951±0.0154**	0.8769±0.0250	0.8816±0.0188	**0.8822±0.0204**	0.8429±0.0352	0.8404±0.0514
M	**EfficientNetB0**	0.8684±0.0434	**0.9239**	0.8681±0.0359	0.8828	0.8221±0.0508	0.8319
	EfficientNetV2B0	0.8248±0.0307	0.8077	0.8250±0.0335	0.8448	0.7527±0.0225	0.7128
	EfficientNetV2B0-21k	0.8937±0.0423	0.8975	0.9077±0.0451	**0.9294**	0.7837±0.0358	0.7500
	ResNetV1-50	**0.9287±0.0234**	0.8960	0.9008±0.0380	0.8879	**0.8456±0.0227**	**0.8542**
	ResNetV2-50	0.9179±0.0304	0.9088	0.9018±0.0337	0.8685	0.8354±0.0240	0.8076
	MobileNetV1	0.9184±0.0230	0.8878	**0.9208±0.0292**	0.9123	0.8146±0.0265	0.7884
	MobileNetV2	0.8660±0.0389	0.8515	0.8702±0.0410	0.8677	0.7745±0.0297	0.8139
	Average (mean±std)	**0.8883±0.0373**	0.8819±0.0396	0.8849±0.0327	**0.8848±0.0286**	0.8041±0.0343	0.7941±0.0487
	Average (mean±std)	**0.8917±0.0277**	0.8794±0.0319	0.8833±0.0257	**0.8835±0.0239**	0.8235±0.0390	0.8172±0.0537

Table 4.2: Balanced Accuracy Results of Seven CNN Models trained in the Original, Reinhard and Macenko FBCG Datasets.

Models	Balanced Accuracy					
	Original		Reinhard		Macenko	
	CV (mean±std)	Test	Mean CV (mean±std)	Mean Test (mean±std)	Mean CV (mean±std)	Mean Test (mean±std)
EfficientNetB0	0.9303±0.0322	0.9518	0.8560±0.0313	0.8703±0.0389	0.8529±0.0266	**0.8795±0.0461**
EfficientNetV2B0	0.9076±0.0398	0.9024	0.8378±0.0555	0.8490±0.0576	0.8008±0.0417	0.7884±0.0681
EfficientNetV2B0-21k	**0.9666±0.0185**	**0.9524**	**0.9058±0.0196**	0.8353±0.0558	0.8617±0.0679	0.8590±0.0957
ResNetV1-50	0.9253±0.0310	0.9239	0.8818±0.0241	**0.9029±0.0074**	0.8917±0.0423	0.8794±0.0222
ResNetV2-50	0.9346±0.0156	0.9198	0.8753±0.0258	0.8950±0.0083	0.8850±0.0437	0.8616±0.0509
MobileNetV1	0.9518±0.0232	**0.9524**	0.8765±0.0266	0.8446±0.0186	0.8846±0.0606	0.8628±0.0656
MobileNetV2	0.9362±0.0322	0.9128	0.8794±0.0114	0.8682±0.0197	0.8369±0.0541	0.8444±0.0276
Average (mean±std)	**0.9361±0.0189**	**0.9308±0.0211**	0.8732±0.0214	0.8665±0.0255	0.8591±0.0325	0.8536±0.0312

Table 4.3: Macro Precision Results of Seven CNN Models trained in Reinhard (denoted as R) and Macenko (denoted as M) FBCG Datasets with Three Different Templates (denoted as T1, T2 and T3).

SN	**Models**	**Macro Precision**					
		T1		**T2**		**T3**	
		CV (mean±std)	**Test**	**CV (mean±std)**	**Test**	**CV (mean±std)**	**Test**
R	**EfficientNetB0**	0.8679±0.0348	0.8879	0.8509±0.0493	0.8865	0.7937±0.0479	0.7941
	EfficientNetV2B0	0.8856±0.0328	0.8750	0.8627±0.0382	0.8953	0.7790±0.0390	0.7452
	EfficientNetV2B0-21k	**0.9144±0.0385**	0.8937	**0.8960±0.0491**	0.8406	**0.8805±0.0281**	0.7892
	ResNetV1-50	0.9010±0.0235	**0.8951**	0.8892±0.0468	**0.9117**	0.8471±0.0677	0.8768
	ResNetV2-50	0.8882±0.0167	0.8869	0.8870±0.0419	0.8898	0.8589±0.0418	**0.9012**
	MobileNetV1	0.8977±0.0317	0.8410	0.8877±0.0642	0.8388	0.8459±0.0670	0.8740
	MobileNetV2	0.8668±0.0493	0.8238	0.8785±0.0490	0.8873	0.8443±0.0541	0.8125
	Average (mean±std)	**0.8888±0.0174**	0.8719±0.0282	0.8789±0.0163	**0.8786±0.0279**	0.8356±0.0361	0.8276±0.0572

Table 4.3 (Continued)

M	**EfficientNetB0**	0.8516±0.0237	0.8845	0.8196±0.0355	0.8333	0.7668±0.0488	0.7813
	EfficientNetV2B0	0.8172±0.0299	0.8056	0.8185±0.0253	0.8425	0.6976±0.0499	0.6117
	EfficientNetV2B0-21k	0.8809±0.0265	0.8873	0.8997±0.0414	0.8955	0.7364±0.0361	0.7009
	ResNetV1-50	**0.9233±0.0279**	0.8871	0.8940±0.0466	0.8980	0.8277±0.0279	**0.8223**
	ResNetV2-50	0.9145±0.0226	**0.9063**	0.8923±0.0327	0.8800	0.8047±0.0133	0.7619
	MobileNetV1	0.8943±0.0340	0.8787	**0.9253±0.0425**	**0.9041**	0.7963±0.0463	0.7410
	MobileNetV2	0.8367±0.0425	0.8188	0.8291±0.0518	0.8506	0.7457±0.0223	0.7785
	Average (mean±std)	**0.8741±0.0401**	0.8669±0.0385	0.8684±0.0445	**0.8720±0.0293**	0.7679±0.0450	0.7425±0.0688
Average (mean±std)		**0.8814±0.0307**	0.8694±0.0325	0.8736±0.0326	**0.8753±0.0277**	0.8018±0.0527	0.7850±0.0751

Table 4.4: Macro Precision Results of Seven CNN Models trained in the Original, Reinhard and Macenko FBCG Datasets.

Models	**Macro Precision**					
	Original		**Reinhard**		**Macenko**	
	CV (mean±std)	**Test**	**Mean CV (mean±std)**	**Mean test (mean±std)**	**Mean CV (mean±std)**	**Mean test (mean±std)**
EfficientNetB0	0.9161±0.0408	0.9511	0.8375±0.0389	0.8562±0.0538	0.8127±0.0428	0.8330±0.0516
EfficientNetV2B0	0.8988±0.0429	0.9046	0.8424±0.0561	0.8385±0.0814	0.7778±0.0694	0.7533±0.1240
EfficientNetV2B0-21k	**0.9646±0.0174**	0.9524	**0.8970±0.0170**	0.8412±0.0523	0.8390±0.0894	0.8279±0.1101
ResNetV1-50	0.9244±0.0358	0.9169	0.8791±0.0283	**0.8945±0.0175**	**0.8817±0.0490**	**0.8691±0.0409**
ResNetV2-50	0.9199±0.0276	0.9012	0.8780±0.0166	0.8926±0.0076	0.8705±0.0581	0.8494±0.0769
MobileNetV1	0.9526±0.0180	**0.9545**	0.8771±0.0275	0.8513±0.0197	0.8720±0.0673	0.8413±0.0878
MobileNetV2	0.9339±0.0251	0.9028	0.8632±0.0174	0.8412±0.0403	0.8038±0.0505	0.8160±0.0361
Average (mean±std)	**0.9300±0.0224**	**0.9262±0.0253**	0.8678±0.0214	0.8594±0.0242	0.8368±0.0399	0.8271±0.0367

Table 4.5: Macro F1-Score Results of Seven CNN Models trained in Reinhard (denoted as R) and Macenko (denoted as M) FBCG Datasets with Three Different Templates (denoted as T1, T2 and T3).

SN	**Models**	**Macro F1-score**					
		T1		**T2**		**T3**	
		CV (mean±std)	**Test**	**CV (mean±std)**	**Test**	**CV (mean±std)**	**Test**
R	**EfficientNetB0**	0.8693±0.0311	**0.8925**	0.8553±0.0539	0.8852	0.7963±0.0573	0.8082
	EfficientNetV2B0	0.8798±0.0311	0.8740	0.8448±0.0513	0.8899	0.7668±0.0512	0.7540
	EfficientNetV2B0-21k	**0.9155±0.0383**	0.8790	**0.9010±0.0423**	0.8400	**0.8748±0.0387**	0.7604
	ResNetV1-50	0.8979±0.0217	0.8920	0.8875±0.0461	**0.9113**	0.8478±0.0632	0.8793
	ResNetV2-50	0.8886±0.0217	0.8911	0.8794±0.0334	0.8860	0.8432±0.0633	**0.8992**
	MobileNetV1	0.8965±0.0289	0.8319	0.8736±0.0754	0.8494	0.8458±0.0669	0.8874
	MobileNetV2	0.8684±0.0459	0.8363	0.8778±0.0535	0.8848	0.8512±0.0475	0.8163
	Average (mean±std)	**0.8880±0.0170**	0.8710±0.0262	0.8742±0.0190	**0.8781±0.0247**	0.8323±0.0372	0.8293±0.0603

Table 4.5 (Continued)

M	**EfficientNetB0**	0.8559±0.0342	0.9013	0.8405±0.0346	0.8552	0.7841±0.0527	0.8039
	EfficientNetV2B0	0.8147±0.0214	0.7947	0.8128±0.0281	0.8377	0.7113±0.0350	0.6418
	EfficientNetV2B0-21k	0.8841±0.0352	0.8920	0.9001±0.0422	**0.9117**	0.7332±0.0240	0.7076
	ResNetV1-50	**0.9244±0.0253**	0.8891	0.8928±0.0444	0.8868	**0.8294±0.0132**	**0.8359**
	ResNetV2-50	0.9143±0.0234	**0.9060**	0.8938±0.0339	0.8690	0.8085±0.0184	0.7825
	MobileNetV1	0.9020±0.0282	0.8785	**0.9212±0.0343**	0.9044	0.7934±0.0343	0.7589
	MobileNetV2	0.8494±0.0408	0.8319	0.8429±0.0482	0.8557	0.7419±0.0252	0.7883
	Average (mean±std)	**0.8778±0.0396**	0.8705 ±0.0415	0.8720±0.0397	**0.8744 ±0.0275**	0.7717±0.0434	0.7598 ±0.0655
Average (mean±std)		**0.8829±0.0297**	0.8707±0.0333	0.8731±0.0299	**0.8762±0.0252**	0.8020±0.0500	0.7946±0.0704

Table 4.6: Macro F1-Score Results of Seven CNN Models trained in the Original, Reinhard and Macenko FBCG Datasets.

Models	**Macro F1-score**					
	Original		**Reinhard**		**Macenko**	
	CV **(mean±std)**	**Test**	**Mean CV** **(mean±std)**	**Mean test** **(mean±std)**	**Mean CV** **(mean±std)**	**Mean test** **(mean±std)**
EfficientNetB0	0.9211±0.0378	0.9494	0.8403±0.0387	0.8620±0.0467	0.8268±0.0378	0.8535±0.0487
EfficientNetV2B0	0.9000±0.0416	0.8982	0.8305±0.0578	0.8393±0.0743	0.7796±0.0592	0.7581±0.1030
EfficientNetV2B0-21k	**0.9642±0.0184**	0.9484	**0.8971±0.0206**	0.8265±0.0604	0.8391±0.0921	0.8371±0.1126
ResNetV1-50	0.9206±0.0334	0.9175	0.8777±0.0264	0.8942±0.0161	**0.8822±0.0484**	**0.8706±0.0301**
ResNetV2-50	0.9259±0.0202	0.9096	0.8704±0.0240	0.8921±0.0067	0.8722±0.0561	0.8525±0.0634
MobileNetV1	0.9506±0.0214	**0.9487**	0.8720±0.0254	0.8562±0.0284	0.8722±0.0689	0.8473±0.0776
MobileNetV2	0.9314±0.0305	0.9058	0.8658±0.0135	0.8458±0.0352	0.8114±0.0603	0.8253±0.0342
Average **(mean±std)**	**0.9305±0.0211**	**0.9254±0.0227**	0.8648±0.0226	0.8594±0.0257	0.8405±0.0376	0.8349±0.0367

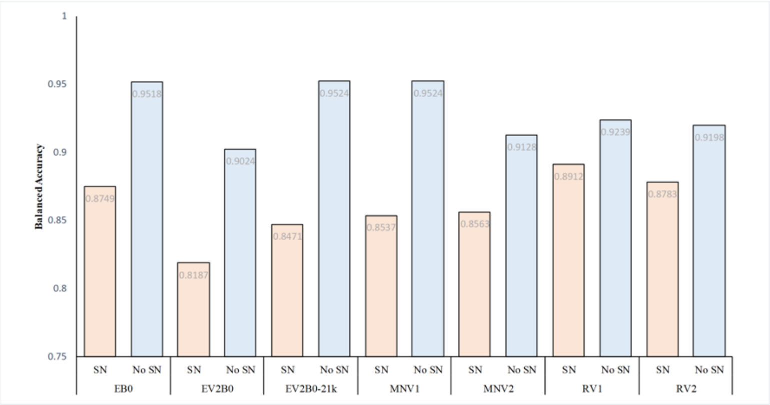


Figure 4.1: Comparing Balanced Accuracy of the CNN Models with SN and without SN across 7 CNN Architectures (EfficientNetB0, EfficientNetV2B0, EfficientNetV2B0-21k, MobileNetV1, MobileNetV2, ResNetV1-50, and ResNetV2-50) on Test Set.

The value in the SN bar represents the average balanced accuracy of the CNN models generated from Reinhard and Macenko normalised images. The graph demonstrates that CNN models without SN perform consistently better on all the tested CNN architectures.

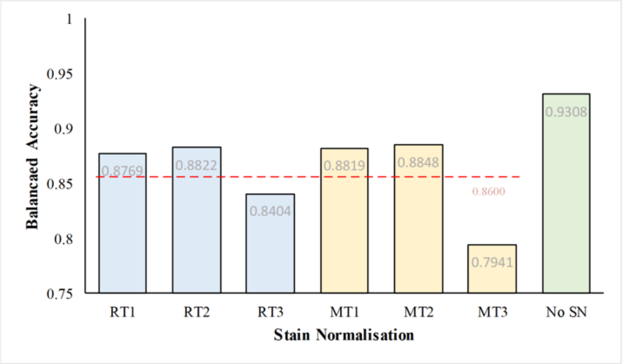


Figure 4.2: Comparison of Balanced Accuracy (on Test Images) of IDC Grading Classification using Images pre-processed by: 1) Reinhard Technique (in blue) using Templates 1, 2, 3 (denoted by RT1, RT2, RT3, respectively); 2) Macenko Technique (in gold) using Template 1, 2, 3 (denoted by MT1, MT2, and MT3, respectively); and 3) without SN.

The value in the bar represents the average value of balanced accuracy across seven selected CNN models. The exact balanced accuracy for each model on test images can be found in Table 4.1 and Table 4.2. The red dashes represent the average balanced accuracy obtained from Reinhard and Macenko techniques. This graph demonstrates the key findings that the IDC grading classification performance using images without SN outperforms those cnn models using Reinhard and Macenko techniques. This finding is consistent with other metrics such as macro precision and macro F1-score (referring to Table 4.3, Table 4.4, Table 4.5, and Table 4.6 respectively). This key finding is opposed to the common practice of applying both SN in IDC grading classification and thus is impactful to guiding future design of automated IDC grading classification systems. The underlying cause for this finding, however, remains unknown.

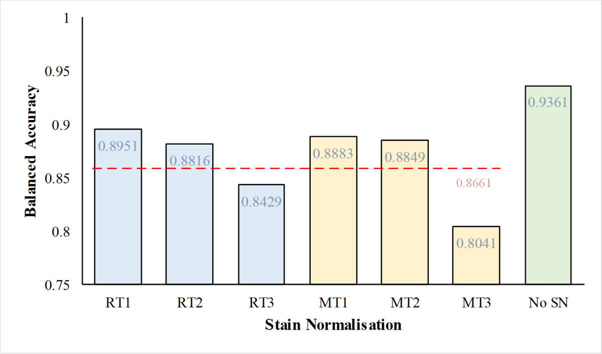


Figure 4.3: Five-Fold CV on the Training Set. The Value in the Bar represents the Average Balanced Accuracy (Five-Fold Cross-Validation) across Seven Selected CNN Models in Grading Classification.

The red dashes represent the average balanced accuracy obtained from Reinhard and Macenko Techniques. The details of the result of CV can be found in Table 4.1. The graph demonstrates that the result from CV aligns with the key findings from Figure 4.1 (IDC grading classification using images without SN outperforms IDC grading classification with images stained using Reinhard and Macenko techniques), implying that the result of the critical finding is reliable and that the result remains valid on a different combination of test images.

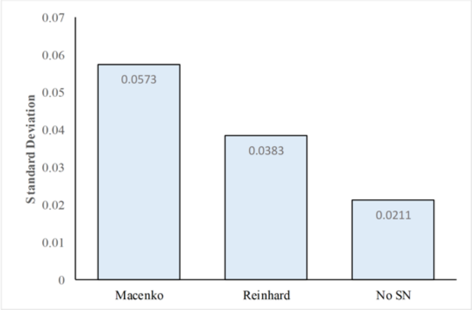
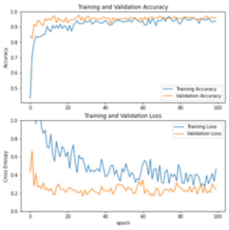
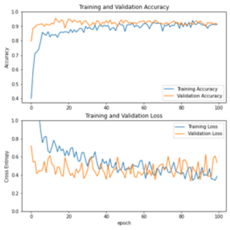


Figure 4.4: Standard Deviation of Balanced Accuracy (on Test Images) across Seven Selected CNN Models by (1) Reinhard Technique, (2) Macenko Technique, and (3) without using SN.

It is noted that the value in the bar of Reinhard or Macenko represents the standard deviation for 27 balanced accuracies (produced by grading classification's balanced accuracy on test images using seven CNNs repeated for three templates). The result shows that the standard deviation for category "No SN" is the lowest, indicating that models trained without SN give the highest stability.

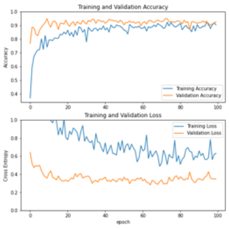
Original **Best** **Worst**

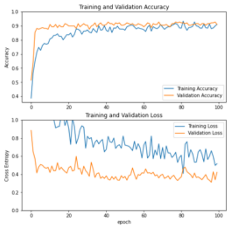




EfficienNetV2B0-21k EfficientNetV2B0

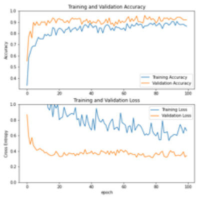
R **T1**

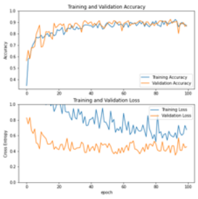




ResNetV1-50 MobileNetV1

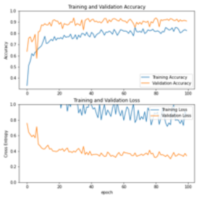
T2

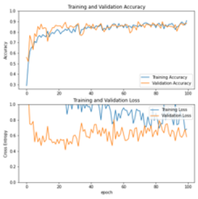




ResNetV1-50 EfficienNetV2B0-21k

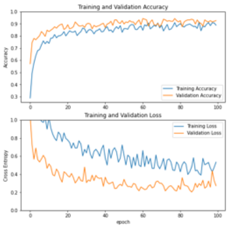
T3

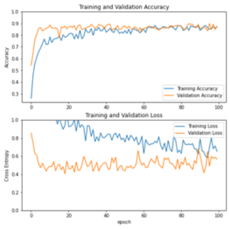




ResNetV2-50 EfficienNetV2B0-21k

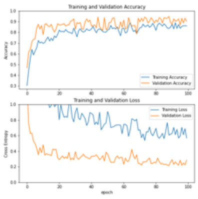
M **T1**

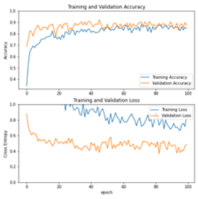




EfficientNetB0 EfficientNetV2B0

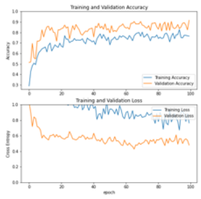
T2

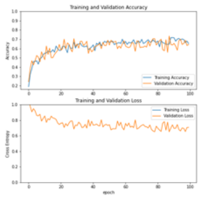




EfficienNetV2B0-21k EfficientNetV2B0

T3



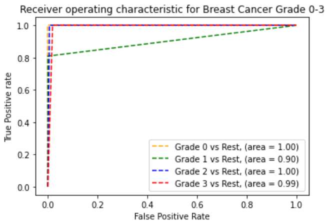


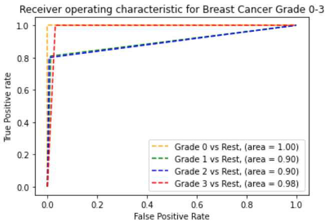
ResNetV1-50 EfficientNetV2B0

Figure 4.5: Accuracy and Loss Curves of the Best and Worst-Performing CNN Models in the Original and SN FBCG Test Sets (20 per cent of the FBCG Dataset).

None of the model curves indicates model overfitting since the validation accuracy curves are higher than the training accuracy curves and the validation loss curves are lower than the training loss curves. The validation loss curves are significantly lower than the training curves are most likely the result of the implementation of dropout layers in the CNN model frameworks.

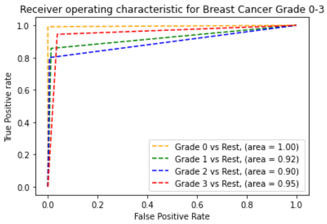
Original **Best** **Worst**

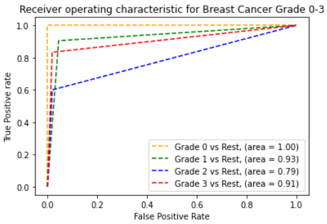




EfficienNetV2B0-21k EfficientNetV2B0

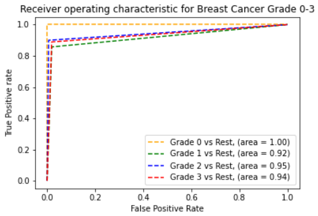
R **T1**

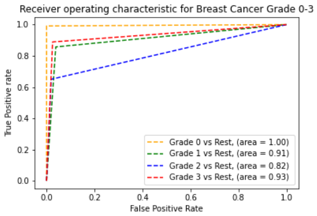




ResNetV1-50 MobileNetV1

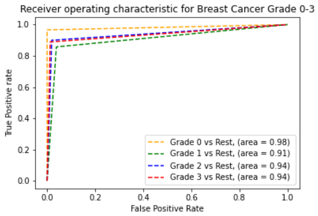
T2

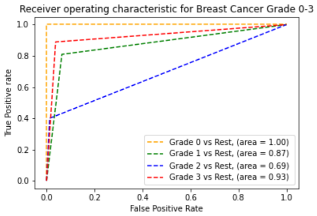




ResNetV1-50 EfficienNetV2B0-21k

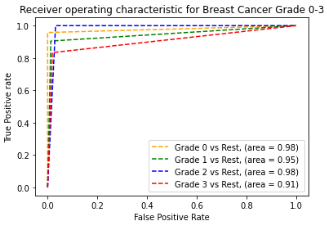
T3

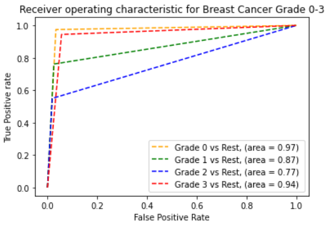




ResNetV2-50 EfficienNetV2B0-21k

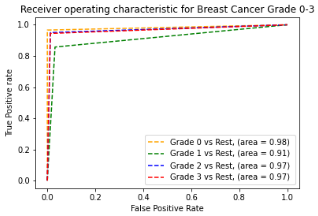
M **T1**

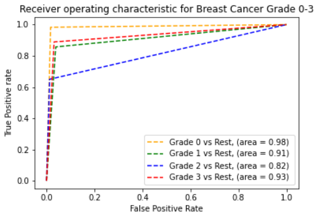




EfficientNetB0 EfficientNetV2B0

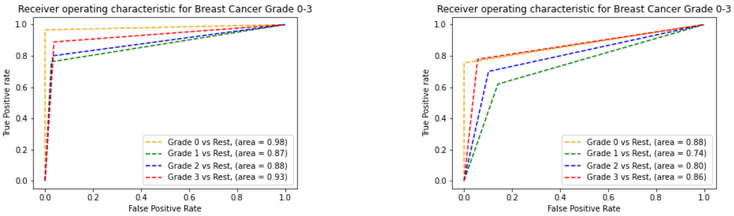
T2





EfficienNetV2B0-21k EfficientNetV2B0

T3

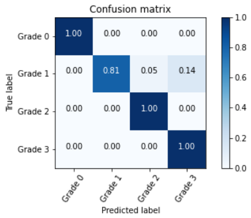


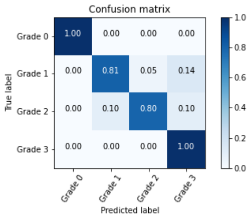
ResNetV1-50 EfficientNetV2B0

Figure 4.6: ROC Curves of the Best and Worst-Performing CNN Models in the Original and SN FBCG Test Sets (20 per cent of the FBCG Dataset).

It shows that, on average, all the chosen CNN models (except Macenko T3 EfficientNetV2-B0) exhibit the highest performance in identifying Grade 0 and the lowest performance in identifying Grade 1. The original EfficientNetV2-B0-21k (top left) scored the top overall AUC scores, whereas the Macenko T3 EfficientNetV2-B0 scored the lowest, supporting that IDC grading classification performance using images without SN outperforms those CNN models using Reinhard and Macenko techniques.

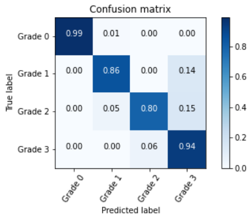
Original **Best** **Worst**

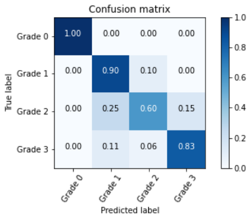




EfficienNetV2B0-21k EfficientNetV2B0

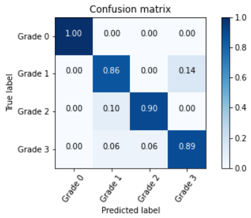
R **T1**

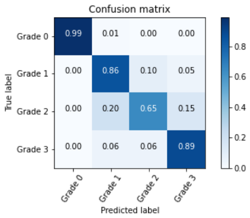




ResNetV1-50 MobileNetV1

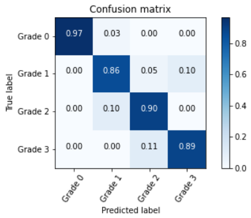
T2

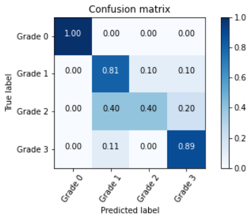




ResNetV1-50 EfficienNetV2B0-21k

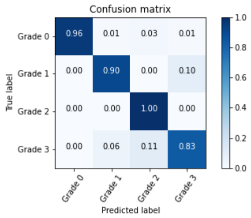
T3

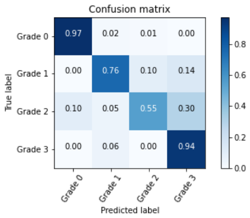




ResNetV2-50 EfficienNetV2B0-21k

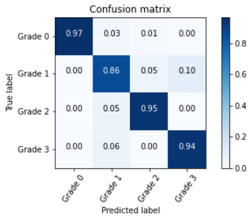
M **T1**

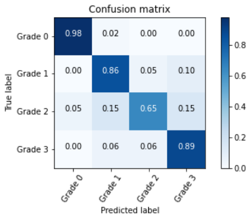




EfficientNetB0 EfficientNetV2B0

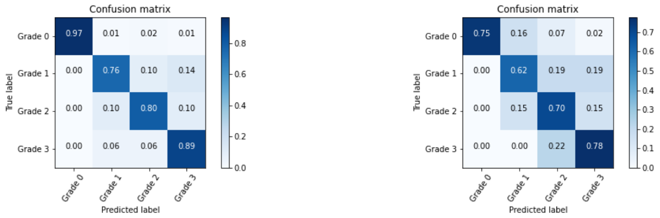
T2





EfficienNetV2B0-21k EfficientNetV2B0

T3

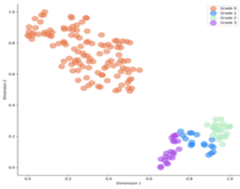


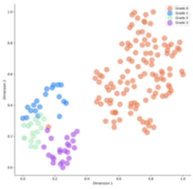
ResNetV1-50 EfficientNetV2B0

Figure 4.7: Normalised Confusion Matrices of the Best and Worst-Performing CNN Models in the Original and SN FBCG Test Sets (20 per cent of the FBCG Dataset).

The confusion matrices indicate that all the chosen CNN models (except Macenko T3 EfficientNetV2-B0) exhibit the highest performance in identifying Grade 0 and the lowest in identifying Grade 1.

Original **Best** **Worst**

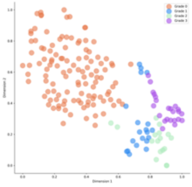


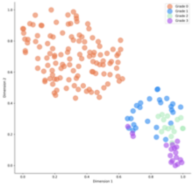


EfficienNetV2B0-21k

EfficientNetV2B0

R **T1**

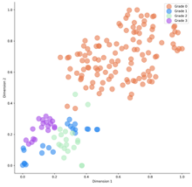


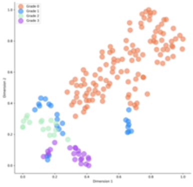


ResNetV1-50

MobileNetV1

T2

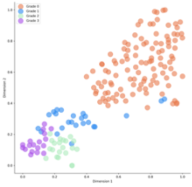


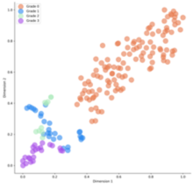


ResNetV1-50

EfficienNetV2B0-21k

T3

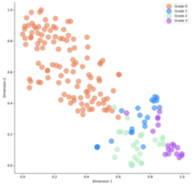


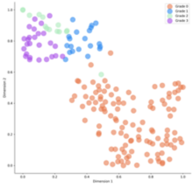


ResNetV2-50

EfficienNetV2B0-21k

M **T1**

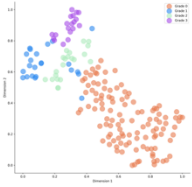


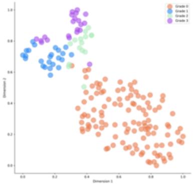


EfficientNetB0

EfficientNetV2B0

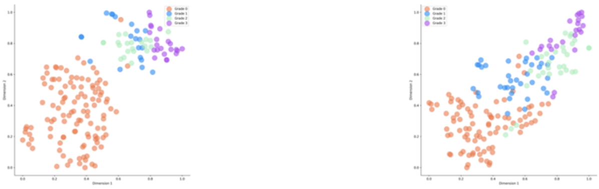
T2





EfficienNetV2B0-21k EfficientNetV2B0

T3



ResNetV1-50 EfficientNetV2B0

Figure 4.8: Visual Comparison of the Distribution of Learned Features t-SNE of the Best And Worst-Performing CNN Models in the Original and SN FBCG Test Sets.

The test set is shown in colours for different classes; orange refers to Grade 0, blue refers to Grade 1, green refers to Grade 2, and purple refers to Grade 3. The figure shows that all selected CNN models can separate Grade 0 well from other grades. The figure also depicts that SN has no distinct effect on the separation of the classes.

4.3 Limitation of Study

In this study, the employed dataset was inspired by Abdelli, et al. (2020). As a result, the experimental results are applicable to the FBCG dataset and comparable to the work of Abdelli, et al. (2020) only. This study examined seven state-of-the-art CNN architectures (EfficientNetB0, EfficientNetV2B0-21k, ResNetV1-50, ResNetV2-50, MobileNetV1, and MobileNetV2); additional CNN architectures were omitted due to time constraints and limited resources. The methodology involved end-to-end feature extraction via transfer learning using pre-trained CNN architectures from TensorFlow Hub. Nonetheless, this study omitted fine-tuning of pre-trained CNN architecture. If fine-tuning is performed in the correct location within the model architecture, it may improve the performance of CNNs without inducing overfitting. Two conventional SN techniques (Reinhard and Macenko) were included in the CNN pre-processing pipeline via StainTools (Byfield, 2020). This study omitted deep learning-based SN techniques. Hence, the study outcome is only applicable to similar studies that employed Reinhard and Macenko techniques in histopathological classification tasks. This study only explored three different selection conditions for template images. Hence, these images may not fully represent the colour characteristics of the whole FBCG dataset.

4.4 Challenges of Study

Among the challenges encountered in this study is the issue of overfitting. The adopted FBCG dataset is relatively small compared to other histopathological breast cancer datasets (BreaKHis). As a result, overfitting may occur when training with more complex CNN architectures. Hence, several augmentation layers (random flip, random rotation, and random zoom) were included in the CNN pipeline to augment the dataset while generating more inputs. Furthermore, two dropout layers were included in the CNN framework to randomly nullify input units at a specified rate during model training, mitigating the risk of overfitting. Working with an unbalanced dataset is another of the hardships encountered in this study. As a result, the CNN model was prone to predict the majority class. Thus, the class weighting technique was applied by giving the minority class a higher weight in the model cost function to impose a more significant penalty on the minority class.

4.5 Summary

In this study, seven pre-trained CNN architectures: (1) EfficientNetB0, (2) EfficientNetV2B0, (3) EfficientNetV2B0-21k, (4) ResNetV1-50, (5) ResNetV2-50, (6) MobileNetV1, and (7) MobileNetV2 were adopted to classify the original, Reinhard-normalised and Macenko-normalised FBCG datasets with three different template images. The findings indicate that templates T1 and T2 outperformed T3 in the IDC grading task, suggesting that selecting the right template image may not necessarily improve the performance of CNN models if ineffective SN is employed. Furthermore, the results show that CNN models trained in the original FBCG dataset outperformed both SN techniques (see Figure 4.1, Figure 4.2 and Figure 4.3) with the highest stability (see Figure 4.4). Our result challenged the general presumption that SN is essential for top performance in histopathological classification tasks. Thus, SN may be unnecessary to be included in the CNN pre-processing step to improve CNN performance if the effective CNN architecture is employed.

The limitations of the study include the adoption of (1) one specific dataset (FBCG dataset), (2) seven pre-trained CNN architectures, (3) end-to-end feature extraction via transfer learning, (4) two SN techniques (Reinhard and Macenko) via StainTools (), and (5) three template images only. The main challenges faced in this study are the risk of overfitting and an unbalanced dataset. Therefore, augmentation and dropout layers and the class weighting technique are utilised, attempting to overcome these challenges.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

This study investigated the importance of Reinhard and Macenko SN techniques in IDC grading with histopathological images using CNN. Seven pre-trained CNN architectures were employed to classify the Four-Breast-Cancer-Grades (FBCG) dataset into four classes via transfer learning: Grade 0, Grade 1, Grade 2, and Grade 3. Based on the experimental results, CNN models trained in the original FBCG dataset outperformed both SN techniques, contesting the general presumption that SN is vital to achieving top performance in the histopathological classification tasks. Comparing the two SN techniques, Reinhard average scores outperformed the Macenko across all evaluation metrics in cv and test results while being more consistent in performance. Therefore, this study suggests that SN is not considered a necessary step to be included in the CNN pre-processing step to improve CNN performance, given that the effective CNN architecture is used. Finally, the outcome of three template images suggests that selecting the right template image may not necessarily enhance the performance of CNN models if ineffective SN is employed.

5.2 Recommendations for future work

In our future development, our study may expand to employ deep learning-based SN techniques in CNN pre-processing steps to compare the performance against original (non-normalised) datasets to support our findings further. Furthermore, this study may consider more recent state-of-the-art CNN architectures and other breast cancer histopathological datasets to conduct our study. Based on the results, it is hypothesised that SN techniques may strip distinct colour features in each IDC grade, leading to poorer CNN performance. Therefore, our study may consider conducting an ablation study regarding the impact of colour features in IDC histopathological images to validate our claim in the future.

REFERENCES

Abdelli, A., Saouli, R., Djemal, K. and Youkana, I., 2020. *Combined Datasets for Breast Cancer Grading Based on Multi-CNN Architectures*. In: 2020 10th International Conference on Image Processing Theory, Tools and Applications, IPTA 2020: IEEE.

American Cancer Society, 2019. Breast cancer facts and figures 2019-2020. *Am Cancer Soc*, [e-journal], pp.1–44. https://doi.org/10.1007/174_2016_83.

Analytics Vidhya, 2020. *How To Dealing With Imbalanced Classes in Machine Learning*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2020/10/improve-class-imbalance-class-weights/> [Accessed 4 September 2021].

Araujo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., Polonia, A. and Campilho, A., 2017. Classification of breast cancer histology images using Convolutional Neural Networks. *PLOS ONE*, [e-journal] 12(6), p.e0177544. https://doi.org/10.1371/JOURNAL.PONE.0177544.

Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S.S., Safwan, M., Alex, V., Marami, B., Prastawa, M., Chan, M., Donovan, M., Fernandez, G., Zeineh, J., Kohl, M., Walz, C., Ludwig, F., Braunewell, S., Baust, M., Vu, Q.D., To, M.N.N., Kim, E., Kwak, J.T., Galal, S., Sanchez-Freire, V., Brancati, N., Frucci, M., Riccio, D., Wang, Y., Sun, L., Ma, K., Fang, J., Kone, I., Boulmane, L., Campilho, A., Eloy, C., Polónia, A. and Aguiar, P., 2019. BACH: Grand challenge on breast cancer histology images. *Medical Image Analysis*, [e-journal] 56, pp.122–139. https://doi.org/10.1016/j.media.2019.05.010.

Bándi, P., Geessink, O., Manson, Q., van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B. E., Lee, B., Paeng, K., Zhong, A., Li, Q., Zanjani, F.G., Zinger, S., Fukuta, K., Komura, D., Ovtcharov, V., Cheng, S., Zeng, S., Thagaard, J., Dahl, A.B., Lin, H., Chen, H., Jacobsson, L., Hedlund, M., Çetin, M., Halici, E., Jackson, H., Chen, R., Both, F., Franke, J., Kusters-Vandevelde, H., Vreuls, W., Bult, P., van Ginneken, B., van der Laak, J. and Litjens, G., 2019. From Detection of Individual Metastases to Classification of Lymph Node Status at the Patient Level: The CAMELYON17 Challenge. *IEEE Transactions on Medical Imaging*, [e-journal] 38(2), pp.550–560. https://doi.org/10.1109/TMI.2018.2867350.

Basavanhally, A., Ganesan, S., Feldman, M., Shih, N., Mies, C., Tomaszewski, J. and Madabhushi, A., 2013. Multi-field-of-view framework for distinguishing tumour grade in ER+ breast cancer from entire histopathology slides. *IEEE Transactions on Biomedical Engineering*, [e-journal] 60(8), pp.2089–2099. https://doi.org/10.1109/TBME.2013.2245129.

Bautista, P., Hashimoto, N. and Yagi, Y., 2014. Colour standardization in whole slide imaging using a colour calibration slide. *Journal of Pathology Informatics*, [e-journal] 5(1). https://doi.org/10.4103/2153-3539.126153.

Bex, T., 2021. *Comprehensive Guide to Multiclass Classification Metrics*. [online] Towards Data Science. Available at: <https://towardsdatascience.com/comprehensive-guide-on-multiclass-classification-metrics-af94cfb83fbd> [Accessed 4 September 2021].

Bolhasani, H., Amjadi, E., Tabatabaeian, M. and Jassbi, S.J., 2020. A histopathological image dataset for grading breast invasive ductal carcinomas. *Informatics in Medicine Unlocked*, [e-journal] 19, p.100341. https://doi.org/10.1016/j.imu.2020.100341.

Byfield, P., 2020. *StainTools*. [online] Available at: <https://github.com/Peter554/StainTools> [Accessed 3 March 2022].

Ciompi, F., Geessink, O., Bejnordi, B.E., de Souza, G.S., Baidoshvili, A., Litjens, G., van Ginneken, B., Nagtegaal, I. and van der Laak, J., 2017. *The importance of stain normalization in colorectal tissue classification with convolutional networks*. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). IEEE.

Dalle, J.-R., Leow, W.K., Racoceanu, D., Tutac, A.E. and Putti, T.C., 2008. *Automatic breast cancer grading of histopathological images*. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference, 2008(2). IEEE.

Dimitropoulos, K., Barmpoutis, P., Zioga, C., Kamas, A., Patsiaoura, K. and Grammalidis, N., 2017. Grading of invasive breast carcinoma through Grassmannian VLAD encoding. *PLoS ONE*, [e-journal] 12(9), p.e0185110. https://doi.org/10.1371/journal.pone.0185110.

Doyle, S., Agner, S., Madabhushi, A., Feldman, M. and Tomaszewski, J., 2008. *Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features*. In: 2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Proceedings*, ISBI*, IEEE.

Bejnordi, B.E., Litjens, G., Timofeeva, N., Otte-Holler, I., Homeyer, A., Karssemeijer, N. and van der Laak, J.A., 2016. Stain Specific Standardization of Whole-Slide Histopathological Images. *IEEE Transactions on Medical Imaging*, [e-journal] 35(2), pp.404–415. https://doi.org/10.1109/TMI.2015.2476509.

Elmore, J.G., Longton, G.M., Carney, P.A., Geller, B.M., Onega, T., Tosteson, A.N.A., Nelson, H.D., Pepe, M.S., H, A.K., Schnitt, S.J., O'Malley, F.P. and Weaver, D.L., 2015. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *Jama*, [e-journal] 313(11), pp.1122–1132. https://doi.org/10.1001/JAMA.2015.1405.

Frkovic-Grazio, S. and Bracko, M., 2002. Long term prognostic value of Nottingham histological grade and its components in early (pT1n0m0) breast carcinoma. *Journal of Clinical Pathology*, [e-journal] 55(2), pp.88–92. https://doi.org/10.1136/jcp.55.2.88.

Ghaznavi, F., Evans, A., Madabhushi, A. and Feldman, M., 2013. Digital imaging in pathology: Whole-slide imaging and beyond. *Annual Review of Pathology: Mechanisms of Disease*, [e-journal] 8, pp331-359. https://doi.org/10.1146/annurev-pathol-011811-120902.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems*, [e-journal] 27. https://doi.org/ 10.3156/jsoft.29.5_177_2.

Guo, Y., Dong, H., Song, F., Zhu, C. and Liu, J., 2018. *Breast Cancer Histology Image Classification Based on Deep Neural Networks*. In: International conference image analysis and recognition. Springer, Cham.

Gupta, V., Singh, A., Sharma, K. and Bhavsar, A., 2017. *Automated Classification for Breast Cancer Histopathology Images: Is Stain Normalization Important?* Computer Assited and Robotic Endoscopy and Clinical Image-Based Procedures. Springer, Cham.

Hamilton, P.W., Allen, D.C., Watt, P.C.H., Patterson, C.C. And Biggart, J.D., 1987. Classification of normal colorectal mucosa and adenocarcinoma by morphometry. *Histopathology*, [e-journal] 11(9). https://doi.org/10.1111/j.1365-2559.1987.tb01897.x.

He, K., Zhang, X., Ren, S. and Sun, J., 2015. *Deep Residual Learning for Image Recognition*. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

He, K., Zhang, X., Ren, S. and Sun, J., 2016. *Identity Mappings in Deep Residual Network*s. In: European conference on computer vision. Springer, Cham.

He, L., Long, L.R., Antani, S. and Thoma, G.R., 2012. Histology image analysis for carcinoma detection and grading. *Computer Methods and Programs in Biomedicine*, [e-journal] 107(3), pp.538–556. https://doi.org/10.1016/j.cmpb.2011.12.007.

Henson, D.E., Ries, L., Freedman, L.S. and Carriaga, M., 1991. Relationship among outcome, stage of disease, and histologic grade for 22,616 cases of breast cancer. The basis for a prognostic index. *Cancer*, [e-journal] 68(10), pp.2142–2149. https://doi.org/10.1002/10970142(19911115)68:10<2142::AID-CNCR2820681010>3.0.CO;2-D.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H., 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv preprint arXiv*, [online] Available at: <http://arxiv.org/abs/1704.04861> [Accessed 2 September 2021].

Johns Hopkins University, 2021. *Staging and grade - breast pathology | Johns Hopkins Pathology*. [online] Available at: <https://pathology.jhu.edu/breast/staging-grade/> [Accessed 7 July 2021].

Kang, H., Luo, D., Feng, W., Zeng, S., Quan, T., Hu, J. and Liu, X., 2021. StainNet: A Fast and Robust Stain Normalization Network. *Frontiers in Medicine*, [e-journal] 8. https://doi.org/10.3389/fmed.2021.746307.

Kassani, S.H., Kassani, P.H., Wesolowski, M.J., Schneider, K.A. and Deters, R., 2019. *Breast Cancer Diagnosis with Transfer Learning and Global Pooling*. In: 2019 International Conference on Information and Communication Technology Convergence (ICTC). IEEE.

Keras, 2021. *Dropout layer*. [online] Keras. Available at: <https://keras.io/api/layers/regularization_layers/dropout/> [Accessed 3 September 2021].

Khan, A.M., Rajpoot, N., Treanor, D. and Magee, D., 2014. A Nonlinear Mapping Approach to Stain Normalization in Digital Histopathology Images Using Image-Specific Colour Deconvolution. *IEEE Transactions on Biomedical Engineering*, [e-journal] 61(6), pp.1729–1738. https://doi.org/10.1109/TBME.2014.2303294.

Komura, D. and Ishikawa, S., 2018. *Machine Learning Methods for Histopathological Image Analysis*. *Computational and Structural Biotechnology Journal*, [e-journal] 16, pp.34-42. https://doi.org/10.1016/j.csbj.2018.01.001.

Lakshmanan, B., Anand, S. and Jenitha, T., 2019. Stain removal through colour normalization of haematoxylin and eosin images: A review. In: *Journal of Physics: Conference Series*. [e-journal] p.012108. https://doi.org/10.1088/1742-6596/1362/1/012108.

Lei, G., Xia, Y., Zhai, D.-H., Zhang, W., Chen, D. and Wang, D., 2020. StainCNNs: An efficient stain feature learning method. *Neurocomputing*, [e-journal] 406, pp.267–273. https://doi.org/10.1016/j.neucom.2020.04.008.

Li, L., Pan, X., Yang, H., Liu, Z., He, Y., Li, Z., Fan, Y., Cao, Z. and Zhang, L., 2020. Multi-task deep learning for fine-grained classification and grading in breast cancer histopathological images. *Multimedia Tools and Applications*, [e-journal] 79(21–22), pp.14509–14528. https://doi.org/10.1007/S11042-018-6970-9.

Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C. and Thomas, N.E., 2009. *A method for normalizing histology slides for quantitative analysis.* In: IEEE International Symposium on Biomedical Imaging. IEEE.

Mccann, M.T., 2015. *Tools for Automated Histology Image Analysis.* Carnegie Melon University. Available at: < https://kilthub.cmu.edu/articles/thesis/Tools_for_Automated_Histology_Image_Analysis/6723926 > [Accessed 7 March 2022]

McCann, M.T., Ozolek, J.A., Castro, C.A., Parvin, B. and Kovačević, J., 2015. *Automated Histology Analysis: Opportunities for signal processing*. IEEE Signal Processing Magazine, [e-journal] 32(1), pp.78-87. https://doi.org/10.1109/MSP.2014.2346443.

Munien, C. and Viriri, S., 2021. Classification of Hematoxylin and Eosin-Stained Breast Cancer Histology Microscopy Images Using Transfer Learning with EfficientNets. *Computational Intelligence and Neuroscience*, 2021 [e-journal]. https://doi.org/10.1155/2021/5580914.

Naik, S., Doyle, S., Agner, S., Madabhushi, A., Feldman, M. and Tomaszewski, J., 2008. Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. In: *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro. IEEE.*

Nawaz, W., Ahmed, S., Tahir, A. and Khan, H.A., 2018. *Classification Of Breast Cancer Histology Images Using ALEXNET.* In: International conference image analysis and recognition. Springer, Cham.

Pan, S.J. and Yang, Q., 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, [e-journal] 22(10), pp.1345–1359. https://doi.org/10.1109/TKDE.2009.191.

Pan, X., Li, L., Yang, H., Liu, Z., He, Y., Li, Z., Fan, Y., Cao, Z. and Zhang, L., 2020. Multi-task Deep Learning for Fine-Grained Classification/Grading in Breast Cancer Histopathological Images. *Studies in Computational Intelligence*, [e-journal] 810, pp.85–95. https://doi.org/10.1007/978-3-030-04946-1_10.

Pêgo, A.P. and Aguiar, P. de C., 2015. *Bioimaging*. *INEB*. Available at: <http://www.bioimaging2015.ineb.up.pt/index.html> [Accessed 2 September 2021].

Priego-Torres, B.M., Sanchez-Morillo, D., Fernandez-Granero, M.A. and Garcia-Rojo, M., 2020. Automatic segmentation of whole-slide H&E stained breast histopathology images using a deep convolutional neural network architecture. *Expert Systems with Applications*, [e-journal] 151. https://doi.org/10.1016/j.eswa.2020.113387.

Qureshi, H., Sertel, O., Rajpoot, N., Wilson, R. and Gurcan, M., 2008. *Adaptive discriminant wavelet packet transform and local binary patterns for meningioma subtype classification*. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Berlin, Heidelberg.

Rakha, E.A., Reis-Filho, J.S., Baehner, F., Dabbs, D.J., Decker, T., Eusebi, V., Fox, S.B., Ichihara, S., Jacquemier, J., Lakhani, S.R., Palacios, J., Richardson, A.L., Schnitt, S.J., Schmitt, F.C., Tan, P.H., Tse, G.M., Badve, S. and Ellis, I.O., 2010. Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Research*, [e-journal] 12(4). https://doi.org/10.1186/bcr2607.

Ramadan, S.Z., 2020. Methods Used in Computer-Aided Diagnosis for Breast Cancer Detection Using Mammograms: A Review. *Journal of Healthcare Engineering*, 2020, [e-journal]. https://doi.org/10.1155/2020/9162464.

Reinhard, E., Ashikhmin, M., Gooch, B. and Shirley, P., 2001. Colour transfer between images. *IEEE Computer Graphics and Applications*, [e-journal] 21(5), pp.34–41. https://doi.org/10.1109/38.946629.

Roy, S., kumar Jain, A., Lal, S. and Kini, J., 2018. A study about colour normalization methods for histopathology images. *Micron*, [e-journal] 114, pp.42–61. https://doi.org/10.1016/j.micron.2018.07.005.

Roy, S., Lal, S. and Kini, J.R., 2019. Novel colour normalization method for hematoxylin eosin stained histopathology images. *IEEE Access*, [e-journal] 7. https://doi.org/10.1109/ACCESS.2019.2894791.

Ruifrok, A.C. and Johnston, D.A., 2001. Quantification of histochemical staining by colour deconvolution. *Analytical and Quantitative Cytology and Histology*, [online] Available at: < https://helios2.mi.parisdescartes.fr/~lomn/Cours/CV/BME/HistoPatho/Color/PythonColorDeconv/Quantification_of_histochemical_staining.pdf > [Accessed 7 March 2022]

Ruiz, A., Sertel, O., Ujaldon, M., Catalyurek, U., Saltz, J. and Gurcan, M., 2007. *Pathological image analysis using the GPU: Stroma classification for neuroblastoma*. In: 2007 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2007. IEEE.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L. C., 2018. *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. In: Proceedings of the IEEE conference on computer vision and pattern recognition.

Schwartz, A.M., Henson, D.E., Chen, D. and Rajamarthandan, S., 2014. Histologic grade remains a prognostic factor for breast cancer regardless of the number of positive lymph nodes and tumour size: a study of 161 708 cases of breast cancer from the SEER program. *Archives of Pathology and Laboratory Medicine*, [e-journal] 138(8), pp.1048–1052. https://doi.org/10.5858/arpa.2013-0435-OA.

Senousy, Z., Abdelsamea, M.M., Mohamed, M.M. and Gaber, M.M., 2021. 3E-net: Entropy-based elastic ensemble of deep convolutional neural networks for grading of invasive breast carcinoma histopathological microscopic images. *Entropy*, [e-journal] 23(5). https://doi.org/10.3390/E23050620.

Shaban, M.T., Baur, C., Navab, N. and Albarqouni, S., 2019. *Staingan: Stain Style Transfer for Digital Histological Images.* In: 2019 Ieee 16th international symposium on biomedical imaging (Isbi 2019. IEEE.

Sharma, G.N., Dave, R., Sanadya, J., Sharma, P. and Sharma, K.K., 2010. Various types and management of breast cancer: An overview. *Journal of Advanced Pharmaceutical Technology and Research*, [online] Available at: </pmc/articles/PMC3255438/> [Accessed 17 June 2021].

Shea, E.K.H., Koh, V.C.Y. and Tan, P.H., 2020. Invasive breast cancer: Current perspectives and emerging views. *Pathology International*, [e-journal] 70(5), pp.242–252. https://doi.org/10.1111/pin.12910.

Shorten, C. and Khoshgoftaar, T.M., 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, [e-journal] 6(1). https://doi.org/10.1186/S40537-019-0197-0.

Spanhol, F.A., Oliveira, L.S., Petitjean, C. and Heutte, L., 2016. A Dataset for Breast Cancer Histopathological Image Classification. *IEEE Transactions on Biomedical Engineering*, [e-journal] 63(7), pp.1455–1462. https://doi.org/10.1109/TBME.2015.2496264.

Stanisavljevic, M., Anghel, A., Papandreou, N., Andani, S., Pati, P., Rüschoff, J.H., Wild, P., Gabrani, M. and Pozidis, H., 2019. *A Fast and Scalable Pipeline for Stain Normalization of Whole-Slide Images in Histopathology*. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops.

Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A. and 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*, [e-journal] 71(3), pp.209–249. https://doi.org/10.3322/caac.21660.

Tan, M. and Le, Q. v., 2019. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. In: 36th International Conference on Machine Learning, ICML 2019. PMLR.
Tan, M. and Le, Q. v, 2021. EfficientNetV2: Smaller Models and Faster Training. *arXiv preprint arXiv*, [online] Available at: <https://arxiv.org/abs/2104.00298v2>.

Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.-M., Ciompi, F. and van der Laak, J., 2019. Quantifying the effects of data augmentation and stain colour normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, [e-journal] 58, p.101544. https://doi.org/10.1016/j.media.2019.101544.

TensorFlow, 2021. *Transfer learning and fine-tuning* . [online] Tensorflow. Available at: <https://www.tensorflow.org/tutorials/images/transfer_learning> [Accessed 3 September 2021].

TensorFlow, 2022. *TensorFlowHub*. [online] Available at: <https://tfhub.dev/> [Accessed 14 March 2022].

Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A.M., Esposito, I. and Navab, N., 2016. Structure-Preserving Colour Normalization and Sparse Stain Separation for Histological Images. *IEEE Transactions on Medical Imaging*, [e-journal] 35(8). https://doi.org/10.1109/TMI.2016.2529665.

Vesal, S., Ravikumar, N., Davari, A.A., Ellmann, S. and Maier, A., 2018. *Classification of Breast Cancer Histology Images Using Transfer Learning.* In: International conference image analysis and recognition. Springer, Cham.

Veta, M., Heng, Y.J., Stathonikos, N., Bejnordi, B.E., Beca, F., Wollmann, T., Rohr, K., Shah, M.A., Wang, D., Rousson, M., Hedlund, M., Tellez, D., Ciompi, F., Zerhouni, E., Lanyi, D., Viana, M., Kovalev, V., Liauchuk, V., Phoulady, H.A., Qaiser, T., Graham, S., Rajpoot, N., Sjöblom, E., Molin, J., Paeng, K., Hwang, S., Park, S., Jia, Z., Chang, E.I.C., Xu, Y., Beck, A.H., van Diest, P.J. and Pluim, J.P.W., 2019. Predicting breast tumour proliferation from whole-slide images: The TUPAC16 challenge. *Medical Image Analysis*, [e-journal] 54. https://doi.org/10.1016/j.media.2019.02.012.

Veta, M., Pluim, J.P.W., van Diest, P.J. and Viergever, M.A., 2014. Breast cancer histopathology image analysis: a review. *IEEE Transactions on Biomedical Engineering*, [e-journal] 61(5), pp.1400–1411. https://doi.org/10.1109/TBME.2014.2303852.

Vo, D.M., Nguyen, N.Q. and Lee, S.W., 2019. Classification of breast cancer histology images using incremental boosting convolution networks. *Information Sciences*, [e-journal] 482, pp.123–138. https://doi.org/10.1016/J.INS.2018.12.089.

Wan, T., Cao, J., Chen, J. and Qin, Z., 2017. Automated grading of breast cancer histopathology using cascaded ensemble with combination of multi-level image features. *Neurocomputing*, [e-journal] 229, pp.34–44. https://doi.org/10.1016/j.neucom.2016.05.084.
Xu, J. and Dong, X., 2020. *A survey of transfer learning in breast cancer image classification*. In: 2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI). IEEE.

Yan, R., Li, J., Rao, X., Lv, Z., Zheng, C., Dou, J., Wang, X., Ren, F. and Zhang, F., 2020. *NANet: Nuclei-Aware Network for Grading of Breast Cancer in HE Stained Pathological Images*. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE.

Zanjani, F.G., Zinger, S., Bejnordi, B.E., van der Laak, J.A.W.M. and de With, P.H.N., 2018a. *Stain normalization of histopathology images using generative adversarial networks*. In: 2018 IEEE 15th International symposium on biomedical imaging (ISBI 2018). IEEE.

Zanjani, F.G., Zinger, S., de With, P.H.N., Bejnordi, B.E. and van der Laak, J., 2018b. Histopathology Stain-Colour Normalization Using Deep Generative Models. *Medical Imaging with Deep Learning*, (Midl), [online] Available at: <https://openreview.net/pdf?id=SkjdxkhoG> [Accessed 7 March 2022]

Zavareh, P.H., Safayari, A. and Bolhasani, H., 2021. *BCNet: A Deep Convolutional Neural Network for Breast Cancer Grading*. *arXiv preprint arXiv*, [online] Available at: <http://arxiv.org/abs/2107.05037> [Accessed 6 August 2021].

Zheng, Y., Jiang, Z., Zhang, H., Xie, F., Hu, Di., Sun, S., Shi, J. and Xue, C., 2021. Stain Standardization Capsule for Application-Driven Histopathological Image Normalization. *IEEE Journal of Biomedical and Health Informatics*, [e-journal] 25(2), pp.337–347. https://doi.org/10.1109/JBHI.2020.2983206.

Zhou, N., Cai, D., Han, X. and Yao, J., 2019. *Enhanced Cycle-Consistent Generative Adversarial Network for Colour Normalization of H&E Stained Images*. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham.

Zhu, J.Y., Park, T., Isola, P. and Efros, A.A., 2017. *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*. In: Proceedings of the IEEE international conference on computer vision.

Zioga, C., Kamas, A., Patsiaoura, K., Dimitropoulos, K., Barmpoutis, P. and Grammalidis, N., 2017. *Breast carcinoma histological images from the Department of Pathology, "Agios Pavlos" General Hospital of Thessaloniki, Greece*. [online] Zenodo. Available at: <https://zenodo.org/record/834910> [Accessed 4 September 2021]

Zulkifli, H., 2018. *Understanding Learning Rates and How It Improves Performance in Deep Learning | by Hafidz Zulkifli | Towards Data Science*. [online] Towards Data Science. Available at: <https://towardsdatascience.com/understanding-learning-rates-and-how-it-improves-performance-in-deep-learning-d0d4059c1c10> [Accessed 4 September 2021].

APPENDICES

Appendix A: Tables

TableA-1: A list of website links for project source code, employed datasets and pre-trained CNN architectures.

Item	Link
Project Source Code	https://github.com/wingatesv/IDCGradingTask.git https://github.com/wingatesv/StainNormalisationIDCGrading.git
BreaKHis Dataset	https://web.inf.ufpr.br/vri/databases/breast-cancerhistopathological-database-breakhis/
BCG Dataset	https://zenodo.org/record/834910#.WXhxt4jrPcs
EfficientNetB0	https://tfhub.dev/tensorflow/efficientnet/b0/feature-vector/1
EfficientNetV2B0	https://tfhub.dev/google/imagenet/efficientnet_v2_imagenet1k_b0/feature_vector/2
EfficientNetV2B0-21k	https://tfhub.dev/google/imagenet/efficientnet_v2_imagenet21k_b0/feature_vector/2
ResNetV1-50	https://tfhub.dev/google/imagenet/resnet_v1_50/feature_vector/5
ResNetV2-50	https://tfhub.dev/google/imagenet/resnet_v2_50/feature_vector/5
MobileNetV1	https://tfhub.dev/google/imagenet/mobilenet_v1_100_224/feature_vector/5
MobileNetV2	https://tfhub.dev/google/imagenet/mobilenet_v2_100_224/feature_vector/5