Interpretable Melanoma Detection for a Clinical Environment

Elliot Naylor

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

University of South Wales
Faculty of Mathematics and Computing

Date

The candidate confirms that the work submitted is his/her own and that appropriate credit has been given where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

The right of Elliot Naylor to be identified as Author of this work has been asserted by him/her in accordance with the Copyright, Designs and Patents Act 1988.

©Year
University of South Wales
and
Candidate Elliot Naylor

Dedication here.

Acknowledgements

Abstract

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Skin cancer is considered amongst the most severe public health concerns, with mortality rates of 2,353 per 100,000 within the United Kingdom (UK) in 2018[103]. Skin cancers can be categorised between melanoma and non-melanoma, whereas melanoma is the most dangerous because it is unpredictable. When left untreated and after growing sufficiently, it can spread to other regions of the body (known as metastatic melanoma), which once progressed is challenging to treat effectively with a 10% survival over ten years in the US[18]. Furthermore, it is beneficial to catch melanoma early because it is the most easily treatable form of cancer, with 86% of cases being preventable[103]. However, melanoma can remain dormant from anywhere between 6 months to 10 years before maturing and becoming a danger to the patient[103]. Another danger of melanoma is its similarity to non-melanoma skin cancers, such as a mimic called seborrhoeic keratosis (SK), which frequently leads to misdiagnoses[43]. There are features unique to SK called fissures, ridges, and hairpin vessels[62]. Problematically these features require trained specialists to recognise them needing more than ten years of experience to have an accuracy of 86% compared to 62% or 56% (3 to 5 years of experience)[66]. However, because of the cost of training new doctors, there are limited available. Dermatologists primarily treat skin conditions (biopsies) and confirm diagnoses submitted by GPs. General practitioners (GP) are the first to diagnose skin conditions and sometimes have limited experience diagnosing them (especially dermatological features). This project aims to improve the accuracy of GP observations by providing tools for the automatic classification of skin lesions. For the previously mentioned reasons, an automatic system should be cost-effective and advantageous to doctors.

Melanoma is a type of malignant skin cancer that accounts for a significant proportion of cancer-related deaths around the world. In 2018 there were approximately 2,353 per 100,000 deaths in the United Kingdom (UK)[103]. Early detection is critical for improving

the diagnosis and survival of patients. However, existing approaches including clinical examinations and dermoscopy, have limitations in terms of accuracy and cost-effectiveness[97]. Machine learning approaches have beaten dermatologists in terms of accuracy[10]. However, these approaches lack explainability implementing such techniques difficult for clinical environments[34]. One concern is the production of realistic, but incorrect results[38]. Another is the use of parallel processes, which describes the creation of an answer with little to no explanation. In this paper, we propose a combined asymmetry approach using shape, colour, and texture analysis alongside a detailed comparison. The technique itself can be used in conjunction with ABCD rules (Asymmetry, border, colour, and dermoscopic features).

Diagnostic procedures are procedures that are performed on patients in order to diagnose conditions. Regarding the diagnosis of melanoma, many types have been utilised for the detection of melanoma and the most favourable are CASH, ABCD rules, and Total Dermoscopy Score (TDS). The ABCD rules and TDS were commonly used because of their simplicity and effectiveness.

Diagnostic procedures are usually based on the doctors medical experience. For example the use-case of this project is for general practioners (GPs), many of which have likely never seen or attempted to diagnose melanoma, and many of which will not have access to dermoscopes for the analysis of dermoscopic structures. So, in this use-case diagnostic procedures including ABCD rules, and CASH are suitable because of their simplicity. The method used by the NHS is also ABCD rules, which is the reasoning for using this method.

Interestingly dermatologists will utilise dermoscopic features and textures, which are more accurate but require sufficient training for the detection.

Diagnostic procedures are instructions developed by doctors to simplify diagnosing conditions. Various methods have been developed to diagnose skin lesions and have greatly improved GP accuracy within clinical environments[68, 104]. Considering melanoma is the most dangerous skin condition, most procedures were developed specifically for early detection. Some include ABCD rules, 2-point checklist, 7-point checklist, and CASH. The most preferred of these techniques are ABCD rules and CASH because they have a higher sensitivity[104] and ABCD rules are generally the most preferred because it is easy to learn and is rapidly calculated[68]. There are several variations of ABCD rules, but, is originally measured using asymmetry, border, colour, and diameter. Diameter is sometimes replaced with dermatological structures because many features (i.e. blue-black signs, pigment networks, pseudopods, streaks, or milia-like cysts[94]) improve the classification accuracy between melanoma and the mimic seborrheic keratosis[25]. Furthermore, automatically measuring diameter is often difficult because it is dependent on the photo apparatus and the distance from the skin lesion, which is rarely consistent in research. Table 1.1.1 describes each rule in more detail, including a scoring system called total dermoscopy score (TDS), where each rule is assigned a score and combined to reach a result of either malignant, suspicious or benign. The criteria is: [(Asymmetry $\times$ 1.3) + (Border $\times$ 0.1) + (Colour $\times$ 0.5) + (Dermoscopic structure $\times$ 0.5)]. Each rule is calculated using the following

Criteria	Methodology	Score	Weight
Asymmetry	Measuring asymmetry involves first finding the centroid and splitting it twice with a 90-degree axis. Each side is subtracted with its opposite half to measure the asymmetry of shape, colour, and dermoscopic structures. If both sides are asymmetrical then the score is 2, one side asymmetrical is a score of 1, and otherwise, the score is 0.	0 - 2	× 1.3
Border	border is found by finding the centroid and drawing lines through it with a 45-degree angle, splitting the skin lesion into eight segments. Border segments might be irregular with convexity, sharp corners, or edges. Irregular segments are incremented by 1, reaching 8 for each segment.	0 - 8	× 1.3
Colour	The area of the skin lesion is up to 6 colours (white, red, light brown, dark brown, blue-grey, black). The score is increased by 1 for each visible colour, reaching a total of 6.	1 - 6	× 0.5
Dermoscopic Structures	Dermoscopic structures are measured by finding structureless areas, pigment networks, atypical networks, dots, and globules. Each visible structure adds a score of 1, reaching a total of 5.	1 - 5	× 0.5

Table 1.1: Total dermoscopy score (TDS) is a scoring system used with ABCD rules to support clinicians when diagnosing melanoma[25]. Each rule is multiplied by weights and the sum of the combined values is the final score, all together: [(Asymmetry × 1.3) + (Border × 0.1) + (Colour × 0.5) + (Dermoscopic structure × 0.5)].

descriptions in Table 1.1.1 and multiplied by their weight and then added together to reach a final score where [< 4.76 = benign, > 4.76 or < 5.45 = suspicious, > 5.45 = melanoma]. The disadvantage is the subjectivity of GP observations relating to their experience. So, it would be beneficial to automate the techniques using algorithms to standardise results and improve GP accuracy.

1.1.1 Total Dermoscopy Score (TDS)

The Total Dermoscopy Score (TDS) is a defined method for assessing the score of the ABCD criteria. Scores less than 4.75 indicates that the lesion is benign, 4.8 to 5.45 indicate a suspicious lesion and anything higher has a high likelihood of being melanoma. Each rule is assigned a score and a weight which are used to assess the structures. Such scoring systems have been implemented into CAD systems successfully[44].

1.1.2 Explainability

The European General Data Protection Regulation (GDPR and ISO/IEC 27001) came into force in May 2018, making "black box" approaches difficult to utilise in a medical environment. Although this is not a ban, the outcome of the results needs to be re-traceable

[42]. These techniques are consistently relied on within medical imaging and without explainability doctors cannot interpret, modify or learn from the provided diagnoses. As such, when using machine learning algorithms, a trusted form of XAI or CAD system must be utilised.

The reasoning behind retractable results is because society is becoming reliant on "black box" approaches that the results of which cannot be analysed. For example, in scenarios where a machine learning model is trained from mostly common lesion samples with a specific skin colour, it might have a lower accuracy against other rarer variations. As these algorithms are not always retractable and there is a lack of image samples within these groups it is difficult to assess whether these algorithms work effectively in rarer scenarios. Therefore, it is ethically and morally incorrect to utilise such methods within a clinical environment. To avoid malpractice the doctor must be provided with relevant and interpretable results to be considered, but not depended on for a diagnoses.

1.2 Datasets

A dataset is a collection of image samples that are often used to compare and validate performance and accuracy of algorithms described in journals, without implementing each technique needed for the comparisons. Furthermore, machine learning techniques require large amounts of image samples for training and without the same amount of relevant images it would be impossible to reproduce some algorithms. For example, one of which utilises a private dataset of 129,450 images consisting of 2,032 different diseases [10], achieving the highest accuracy in classification so far. As a comparison, the largest public dataset called ISIC 2019 has 25,331 images containing 8 different diseases. This is one of the biggest concerns in medical imaging as comparisons with public datasets are not always made, bottlenecking active research within a topic [83]. For this reason, competing against deep learning models has become less prominent and there has been a greater focus on other topics including image synthesis, explainable AI (XAI) and some CAD systems.

Other datasets including MED-NODE [11] provide macroscopic image samples for comparing and creating techniques without a dermoscope. Macroscopic images could have large variations in noise, angles and lighting conditions where the algorithm might fail. For example, the training process of a machine learning algorithm is specific to the type of data it analyses. If the model is used on an image that has a different variation in lighting and angles the algorithm is likely to fail. As such, rules must be followed by the doctor taking the image to insure accurate results.

1.3 Challenges - Project Scenario

This project has some unique components being that the traditional diagnostic procedure used by the dermatologists are based around touching and feeling through pinch tests.

Dataset	Image Samples	Image Type	Availability
ISIC 2019	13,000	Dermoscopic	Public
BCN20000	19,424	Dermoscopic	Public
HAM10000	11,788	Dermoscopic	Public
NHS	unsure	Dermoscopic	Private
Atlas	80	Dermoscopic	Public
MED-NODE	170	Macroscopic	Public

Table 1.2: List of datasets that are used throughout this project. Some of these datasets share images, making the total amount of unique images 25,581

This might include an analysis of roughness, firmness, bleeding, etc. Other information that is considered are age and a lesions location on a body. These components are rarely considered within a CAD framework. A possibility would be to analyse the lesion based around the ABCD rules to separate the component into a benign or malignant skin lesion. The framework would then request additional information through a pinch tests, age and location to narrow the results down to a type skin condition. This additional information would prove valuable to the dermatologists examining the diagnoses at a later date.

1.3.1 Clinical Environment

In a clinical environment, sessions are often short because of the number of patients being diagnosed at a given time on a range of topics including dermatology. For many general practitioners it might be their first attempt at diagnosing a patient with a skin related issue. As described previously many benign lesions have a range of similar features to dangerous cancers, the most common of which are moles and melanoma. It is unrealistic to assume that each doctor has the relevant experience to correctly diagnose a skin lesion. Diagnoses methods including the ABCD rule were created to find some of the most common and dangerous skin lesions such as melanoma. However, there are over 3000 different skin lesions making it unlikely for a doctor to diagnose each effectively. Therefore, a secondary opinion is required from a dermatologist to improve the accuracy of the diagnoses.

Assuming the dermatologist prioritises undiagnosed lesions they can still be left weeks before being diagnosed. However, the main complication is when a general practitioner incorrectly label a skin lesion, leaving it unchecked by a dermatologist indefinitely. This is because it is often impractical to use their limited time to be looking through a range of already diagnosed skin conditions. This is where a computer aided diagnoses system (CAD) can provide a secondary opinion to the doctor by providing tools that would standardize and improve the accuracy of the diagnostic procedure.

The final diagnoses should be defined by the doctor and not the algorithm. For example, the algorithm should be built to provide relevant data supporting the classification process relating ABCD rules or other smaller features. These should be built to persuade and even

teach the doctor of better diagnoses methods, without a too convincing diagnosis.

1.3.2 Data Samples

There are two different environments in which images of skin lesions have been captured which includes hospitals with access dermatologists and surgeries with general practitioners (GPs). Available equipment varies between these locations and general practitioners will often use digital cameras to capture skin lesions while dermatologists have access to dermocopes and other specialised equipment. Therefore the data samples will have variations between dermoscopic and macroscopic images depending on where they were captured.

One of the underlining problems with using macroscopic images for analysis is that they are less likely to contain smaller features which can improve the accuracy of a diagnosis[**empty citation**]. Furthermore, melanoma can contain 6 different colours and, in a room too bright or too dark these colours and other details can be difficult to interpret, leading to an incorrect diagnosis. Most data samples available are

There was no mention of rules used to take these macroscopic images, meaning they might have large variations in lighting, angles and blurring. This makes it substantially difficult to utilise any illumination or noise removal techniques before using them to train a classification model. The training process is not controlled, and any intervening elements might be considered features, worsening the accuracy. It might be possible to train an algorithm only using the limited dermoscopic examples, which would have improved accuracy [41] compared to utilising macroscopic images.

1.4 Summary

Melanoma is one of the most dangerous skin cancers because it is difficult to differentiate between other skin conditions. For example, diagnosing with the ABCD rules is simple and allows for a fast and accurate diagnoses of the most common types of skin lesions, but, is inefficient at diagnosing some types benign lesions (SK) and rare types of melanoma[21]. Alongside the support of diagnostic procedures and Computer-Aided diagnostic (CAD) systems the accuracy continues to increase.

The goal of this project is to improve the diagnostic accuracy of skin lesions within clinical environments through the development of a CAD system. This system would ideally be capable of producing results that can be retraced and understood by the doctor. One of the central problems with this project is that surgeries lack access to specialised equipment including dermoscopes. This means the system has to be developed for and using macroscopic images. These images could potentially contain infinite variations in lighting directions, colour, shadows, blurring, etc. These will hide structures and alter colour intensities. Images are then less likely to contain reliable and consistent features

for the development of machine learning algorithms. This means there will be a decreased accuracy compared to the analysis of dermoscopic images.

1.4.1 UI Development and visualisation

The user interface (UI) being the method in which algorithms are visualised is arguably the most import section of this project because without the means of presenting algorithms in an explainable way they are of little to no use to healthcare professionals. Providing explanations will improve trust from doctors and patients to these algorithms[13]. This is also crucial for the successful implementation of AI technologies within healthcare settings[92].

Computer-aided diagnostic (CAD) frameworks are a collection of algorithms designed to guide decision-making processes within clinical environments[31]. A paper written by Andre Estava demonstrates a deep convolutional neural network (DCNN) that has comparable accuracy to that of dermatologists, trained using 129,450 clinical images consisting of 2,032 different diseases[10]. DCNN generates a collection of artificial neurons organised into layers, where each neuron receives input from a previous layer to perform a computation. The collection of layers is a network, which (once trained) ultimately measures the relationship between input parameters based on provided data. It is important to note that the accuracy is proportionate to the number of images and data quality for training that network. Unfortunately, these image samples are frequently private and unavailable to many institutions. Without adequate image data to test the capabilities of machine learning models, there is no method for measuring these biases and is therefore unsafe to use within clinical environments. Secondly, these approaches will often produce a parallel diagnosis, meaning that results are not always explainable[57]. There are many valuable techniques, but even the best techniques are inadequate for doctors without catering to interpretability. Other techniques are interpretable by considering diagnostic procedures, such as ABCD rules, many of which are described by Ali[7]. Techniques based on diagnostic procedures can be more easily tested for biases and provide further insight to GPs with the means to learn from and understand results. Techniques include support vector machine (SVM), a supervised machine learning that uses regression analysis to categorise labelled data into two or more groups. The advantage means less data for training is required, and the model is interpretable.

Doctors will often only have access to a patient for a short time before moving to another. CAD frameworks are beneficial because they speed up the process, can improve accuracy[31], and ensure the gathering of relevant data (ABCD rules). Furthermore, it could take days for a second opinion from another doctor, where an automatic system immediately provides it. Automated systems should also provide adequate explanations that can be understood quickly and easily by doctors[57]. One method is to provide visual explanations. Many authors[113, 46, 5] describe different ways to measure ABCD rules, including the asymmetry of skin lesions using bi-fold. Automated versions of the procedure

use the centroid and moments of inertia to fold the skin lesion horizontally and vertically along the centroid. The overhung area on both axes is subtracted from the final score to measure asymmetrical or symmetrical. This technique produces an adequate visualisation that can provide GPs with an interpretable result. There is a range of other examples for ABCD rules, including border[46, 113, 6], colour[90, 99, 46], and dermoscopic structures[58] that use a range of interpretable algorithms that produce interpretable results.

1.4.2 Conclusion

Overall, many advanced machine learning techniques using neural networks lack the interpretability required within clinical environments. Furthermore, public datasets lack rarer skin conditions, making finding biases challenging. Automating the ABCD rules can solve this by using a technique that GPs are familiar with, and by using statistical models to extract relevant features (relating to the ABCD rules). This is followed by summarising rules using Bayesian fusion and calculating the significance of individual features.

1.5 Purpose and Research Questions

In this thesis machine learning algorithms were developed for the automatic classification of skin lesions with consideration for its use within clinical environments. Furthermore, the technqiues need to produce an explainable and meaningful response that can support doctors (dermotologists and general practioners) by providing insights and details that might have been missed. Concerning this, two main research questions:

- Are the produced algorithms accurate enough to be properly utilised in a clincal environment?
- Does the developed technique provide meaningful responses that can be properly utilised within a clinical environemnt?

1.6 Scope and Limitations (Use-Case)

Some limitations were set by the company partner based on where they plan to utilise the developed algorithms. In this case the technique is primarly going to be utilised by GP (general practitioners) that are unlikely to have the specific training and equipment for diagnosing skin lesions.

- The algorithms are developed for the detection of Benign Naevi (BN), Melanoma (MM), Seborrhoeic keratosis (SK), Atypical Naevi (AN), Typical Naevi (TN), Squamous Cell Carcinoma (SCC), Basal Cell Carcinoma (BCC). With a primary focus on SK consdering its difficulty to diagnose.

- Macroscopic images will be analysed instead of dermoscopes. This essentially means a standard camera will be used, which includes variations in lighting which might obscure features in skin lesions.
- Meta-data includes area on body, gender, Date of Birth (DOB), department, Diagnosis. The goal is to utilise this data for automatic detection.
- Techniques must be explainable, so that doctors can recognise incorrect responses.

1.7 Target Group

The work is primarly for general practioners (GPs) because it is usually the source of misdiagnosed skin lesions due to the lack of specific training compared to dermatologists. The goal is to improve the accruacy of techniques using a cheap device for capturing data, furthermore automatic classification should provide a more adequate means for dermotologists to search for cases.

1.8 Aim

- Develop an interpretable CAD framework based on the ABCD rules to diagnose skin lesions automatically. The goal is to utilise statistical models to extract each ABCD rule (asymmetry, border, colour, and Dermoscopic structure). Each rule will be trained using individual SVM models and are combined using Bayesian Fusion.

1.9 Objectives

- Develop and validate skin lesion segmentation and border cut-off approach for improved irregularity detection of ABCD rules using SegNet and LBPC.
- Develop and validate melanoma classification based on the diagnostic procedure ABCD rules (asymmetry, border, colour, and dermoscopic structures) for improved interpretability to doctors using various statistical techniques and SVM models.
- Develop and validate combining ABCD rules for the probabilistic analysis of the most dependent features using Bayesian fusion. This could include meta-data for gender, age, touch, feeling, and location on the body.

1.10 Contributions to knowledge

1. **Developing and validating a novel skin lesion segmentation approach for accurate border cut-off segmentation to improve border irregularity**

analysis using SegNet and LBPC.

SegNet is highly accurate at finding the area for the segmentation of skin lesions but is inaccurate for measuring border irregularities because the border cut-off between skin and skin lesion is insufficient. Border irregularity detection necessitates an accurate cut-off for more reliable results, which SegNet does not provide. LBPC solves this problem by exaggerating the cut-off and improving the accuracy of border irregularity detection. However, the disadvantage of LBPC is its inaccuracy when finding the skin lesion area. By combining SegNet and LBPC, detecting the skin lesion area using SegNet, followed by adjusting the border with LBPC; retaining the accuracy of SegNet while improving the border cut-off accuracy. Experimental testing utilising the PH^2 dataset containing expert segmentation data will determine the benefits of segmentation.

2. **Developing and validating a novel asymmetry analysis approach for improved irregular asymmetry detection in skin lesions using moment-based texture analysis for improved bi-fold analysis and superpixels for improved asymmetry colour comparisons.**

   The disadvantage of asymmetry measuring techniques for skin lesions is rotational moments for creating bi-folds. Current bi-folds solely consider the silhouette of the skin lesion, with no consideration towards colour or texture. Furthermore, recent techniques have measured asymmetrical irregularities based on colour and texture. Producing a bi-fold based on the shape, colour, and texture using moment-based texture analysis should improve the accuracy of asymmetry detection. In addition, utilising superpixels to measure colour asymmetry to avoid merging important features improves accuracy. Both techniques will be validated using the PH^2 asymmetrical score.

3. **Developing and validating a novel interpretable melanoma classifier for improved interpretability of ABCD rules (asymmetry, border, colour, and dermoscopic structures) using feature extraction, support vector machines (SVM), and Bayesian fusion.**

   The disadvantage of many neural network-oriented techniques is their lack of adequate interpretability, making them challenging to utilise in clinical environments. However, ABCD rules (asymmetry, border, colour, and dermoscopic structures) are a diagnostic procedure that most doctors are familiar with; therefore, developing a system automating this procedure is beneficial. Feature extraction techniques aim to separate the data essential for each ABCD rule and train an SVM model from the extracted features. For example, bi-folds measure asymmetry, which can be modified to train an SVM model. Repeating this for border, colour, and dermoscopic structures ensures that each rule is independent. Finally, combining the Bayesian fusion results measures the probabilistic significance between ABCD rules and combines them into

benign or malignant. Techniques will be validated using the PH^2 dataset for testing ABCD rules and ISIC 2018 datasets for diagnosis.

4. **Developing and validating a novel interpretable melanoma classifier with meta-data including age, gender, feeling, and location on the body to improve classification accuracy between melanoma and seborrhoeic keratosis (SK) using Bayesian probability for a modifiable probabilistic analysis.**

   Seborrhoeic keratosis (SK) is a melanoma mimic because it sometimes shares clinical features with melanoma. Moreover, differentiating between the two with entirely image data can lead to inaccuracies. Including meta-data age, gender, feeling, and location on the body should improve accuracy because SK appears more frequently on the head or back of old male patients. Bayesian probability networks are considered highly modifiable and can generate results with incomplete input, meaning meta-data is only inputted when necessary, benefiting doctors and improving the diagnosis. The associated organisation has a vast amount of valuable meta-data alongside image data of skin lesions; a private dataset will be created from these results and used to validate results.

Chapter 2

Literature Review

2.1 Introduction

This chapter reviews statistical and neural network algorithms for the automatic classification of melanoma. Following a discussion on the effectiveness of techniques and whether they are useful within clinical environments.

2.2 Background

Melanoma is a deadly skin cancer that frequently results in the death of patients if develops into metastatic melanoma. This refers to when the cancer has burrowed past the skin and makes its way into blood and internal organs. From this point is it far more difficult to remove.

Melanoma develops from melanocyte cells, which in turn produce melanin resulting in skin pigmentation (brown patch of skin). This means there are visual characteristics of melanoma as it continues to grow. Alongside the necessity to improve the diagnostic accuracy the visual characteristics being ideal for the development of computer vision-based algorithms, this has sparked the creation of algorithms and in turn papers.

When doctors utilize a clinical diagnostic tool they should be capable of rationalising and building explanations based on the data provided from that tool. Currently, many techniques[10] called named 'black box' approaches produce parallel diagnosis that lacks adequate explanations for clinical environments. These provide insufficient information for use within some clinical environments[10]. Instead, it would be beneficial for doctors to follow procedures they are familiar with, such as diagnostic procedures including ABCD rules. The reviewed techniques aim to automate the ABCD rules using various statistical and machine-learning techniques. Many are interpretable and suitable for clinical environments.

Hybrid machine learning techniques are recently gaining traction, an example by Ali combines results from both Gaussian naive Bayes (GNB) and a CNN[6] for border irregularity

detection. The CNN ensures high-accuracy classification by finding the relationship between each component, and the GNB is interpretable. Results are combined using an ensemble approach, making a prediction probability. Such techniques are promising for use within clinical environments.

There is a lack of literature describing adequate visual representations for doctors, and it is understandable as there is still little evidence proving that CAD systems improve doctors decision making-processes[84]. It would be beneficial to create literature describing a catalogue of different visualisations that benefit doctors. Putting all this information together, alongside a questionnaire, might provide further insight into the visualisations that might be most useful to doctors.

2.3 Skin Lesions

Skin lesions refer to a section of skin that has an abnormal appearance or growth compared to the surrounding skin. These are separated into two categories including primary which are abnormal skin conditions at birth consisting of birthmarks and moles. Secondary skin conditions are abnormalities obtained after birth which include a range of diseases, cancers and benign lesions. Many of these skin conditions are caused by a range of factors including age, UV light and genetics.

There are over 3000 known skin disorders in the area of dermatology some of which are so rare that it is unlikely for there to be any relevant images samples for analysis. This section holds a discussion on some of the most commonly found benign lesions and skin cancers that this project is concentrated around, the most important of which being melanoma.

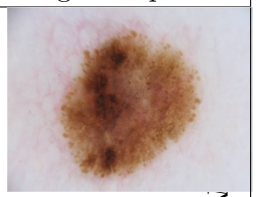
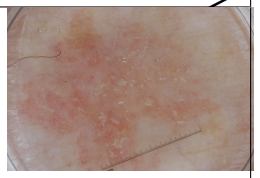
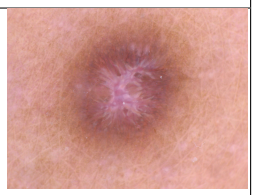
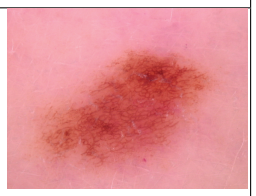
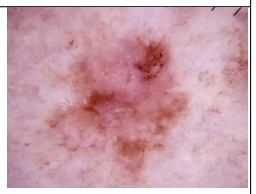
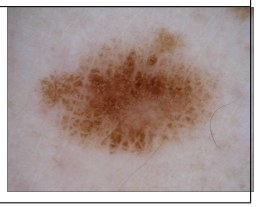
2.3.1 Benign Lesions

Benign lesions are a non-cancerous areas of skin that are safe and unlikely to require any medical treatment. These are common and often share features with dangerous forms of skin cancers including Asymmetry, border and colour. As such, they are frequently given priority over other more serious lesions, slowing down the medical diagnosis procedure. One of the prime examples of this Serborrheic Keratosis (SK) which has similarities to melanoma. The following benign lesions discussed are Melancytic nevus (mole), benign keratosis, dermatofibroma and vascular lesions.

2.3.2 Skin Cancers

Skin cancer is one of the most common cancers that forms in the upper most layers of skin. Malignant cells divide without control spreading from the point of origin. There are many different types of skin cancers under categories of non-melanoma and melanoma. Non-melanoma refers to Basal Cell Carcinoma (BCC) and Squamous Cell Carcinoma, which

are safer because there is considerably less risk of metastasis; spreading to other regions of the body. Regardless skin cancers are easily curable [69] if found at the earliest stages of development. Melanoma however has a higher likelihood of going metastatic, which can only take 6 weeks, making it deadlier than other forms of skin cancer.

Skin Lesion	Type	Description	Image Sample
Melanocytic Nevus	Benign	Melanocytic nevus or a mole is a harmless growth that often appears during childhood, birth or from prolonged exposure to UV light. They appear as pigmented clusters of cells representing a light to dark brown patch of skin that can protrude from the surface.	
Serborrheic Keratosis	Benign	The direct causes of these lesions are unknown, but they become more prevalent as the patient ages so genetics might play a role. These lesions are completely harmless, but the formation of them might be a symptom to a range of other problems.	
Dermatofibroma	Benign	This is a harmless skin lesion that forms within the dermis layer of skin that often related to insect bites or an immune system imbalance. However, being a red tumour like lesion these can be often misplaced with different skin cancers such as desmoplastic melanoma. [23].	
Vascular Lesion	Benign	Vascular lesions are harmless abnormalities otherwise known as birthmarks. These consist of three main categories of haemangiomas, vascular malformations and pyogenic granulomas. Vascular tumours can form which can be benign or malignant.	
Basal Cell Carcinoma	Malignant	Basal cell carcinoma (BCC) is one of the most commonly occurring skin cancer that develops on areas that have suffered from UV damage and develops. These lesions could take years to develop, making it the least dangerous type of skin cancer.	
Squamous Cell Carcinoma	Malignant	Squamous cell carcinoma (SCC) is caused by prolonged exposure to UV radiation and is prominent in people who sunburn easily and are over 70 years old. This type of skin cancer is not the most common but is one of the most devastating form of non-melanoma skin cancers.	
Melanoma	Malignant	Melanoma is the most common and dangerous forms of skin cancers relating to 4% of the population. As melanoma evolves it burrows down through the skin eventually spreading to other regions of the body including lungs, liver, bones, brain, and intestines [27]. This is called metastases, which frequently results in death [86]	

Rules	Description	Example
Asymmetry	Melanoma is typically asymmetrical, meaning that if a line was drawn down the centre and the sides do not match there is a high likelihood that it is melanoma.	Image
Border	Irregular borders are normally a sign of melanoma. Normal moles have a smooth edge all the way around, while this has jagged or notched edges.	Image
Colour	Variations in colour are common in melanoma. Moles typically are a single shade of brown. As it evolves the colour red, white or blue can appear.	Image
Dermoscopic Structures	The diameter is for the analysis of structureless areas, pigment networks, atypical networks, dots and globules. These structures are required for the diagnoses of rarer melanoma and allow for a higher accuracy overall.	Image
Evolution	Any change in size, shape and colour could prove that an area is melanoma. Other factors include bleeding and itching.	Image

Table 2.1: Example of the ABCDE rules with a description and an image. These images were obtained from the ISIC 2019 dataset[102, 24, 26].

2.4 Diagnostic Procedures (ABCD Rules, CASH, 7-Point Checklist, Texture)

Diagnoses procedures were developed to simplify the process and to pursue higher accuracy [80] results from doctors that are not trained specifically in the analysis of skin lesions. Some of these include menzies, 7-point checklist, CASH, ABCD, CHAOS, BLINCK, TADA and pattern analysis. The most popular is ABCD rules because of its combined simplicity and accuracy [**¡survey¿**] when recognising between benign and malignant skin lesions.

The ABCD rules include Asymmetry, Borders, Colours, Diameter/dermoscopic structures and has been expanded to include Evolution[22]. Each feature is assigned a scoring system that can be used to diagnose skin lesions.

The 7-point checklist (7-PCL) is another effective[107] diagnosis method that considers the change in size, irregular pigmentation, irregular border, inflammation, itch/sensation, size (larger than 7mm) and oozing/crusting. This has similarities to the ABCDE rules, but, this method not only uses sight to assess a lesion, but considers the patients sensation and the feel of the lesion. This would be difficult to implement within a framework because the relevant data describing the sensation might not be available.

Another methods called the 3-point Checklist (3-PCL) is a simple diagnoses method that aims to analyse the asymmetry, atypical Network and blue-white structures within a lesion. These features are very specific compared to the other methods and naturally

with the lack of some common features within melanoma it seems likely for this method to be less effective. This method is described for use in a home environment, but atypical networks would be difficult for a non-doctor to recognise.

The Menzies method is a complex technique that has the highest accuracy[22] for diagnosing melanoma using negative and positive features. Negative feature include checking whether the lesion is symmetrical or of a singular colours, either of which define it is non-melanoma. Positive features include blue-white veil, multiple brown dots, pseudopods, radial steaming, scar-like depigmentation, peripheral black globules, five to six colours, mutliple blue-gray dots, broadened network. This method is naturally difficult to utilise within a clinical environment because of the expertise required recognise the features. Another problem is that a diagnoses needs to be fast because of the quantity of patients being processed, but, the number of rules does not support this.

Pattern analysis is a simultaneous analysis of all components within a lesion. By far the most difficult method, requiring time and great expertise to utilise in a clinical environment. This is likely to only ever be used by expert dermatologists that have possibly analysed hundreds of lesions. The method consists of two categories including global features and local features. Global features describe the structure of the lesion which could include multicomponent patterns, unspecific pattern (Structureless or irregular), parallel patterns (Ridges, palms and soles). Local features analyses the components including atypical pigment networks, dots/globules, asymmetrical, irregular streaks, five or six colours, blue-white veil, etc. If the doctor can recognise each component and the correlation between them, this has the potential to be the most effective method for diagnoses.

Another complementary sign is the "ugly duckling" approach, which urges for a comparison between other moles on the patient's body. The one that is unlike any of the others is often suspect for malignancy [45]. This is an interesting method as it requires less prior experience to correctly diagnose the patient. However, there are no direct examples proving the effectiveness in a clinical environment.

The goal of these methods is to simplify the process to improve the performance and accuracy of a diagnosis. However, the rules are not always effective against different types of melanoma. Examples of these are Amelanotic melanoma which is colourless and lacks the pigments required for analysing border and colour [72]. Another is called nodular melanoma which appears darker and lacks in colour variation (black) or dotted features that are often required diagnoses. This is likely why the ABCD rules was modified from diameter to dermoscopic structures, providing more than enough information for an accurate diagnoses.

2.4.1 Computer-aided Diagnosis (CAD)

Most cases are first assigned to a general practitioner (GP), which do not have training in dermatology and skills are more globally oriented. This means they might struggle to recognise some of the skin features needed for a correct diagnoses. In these scenarios a secondary opinion is required, but, there are often limited dermatologists on hand, which can

often lead to a misdiagnoses. To combat this CAD systems were developed to assist these doctors with an automatic secondary opinion; improving the likelihood of more frequently correct diagnoses. The main types of CAD systems are broken down into a smaller criteria including dictionary based features, clinically relevant features and deep learning[16].

Bag-of-Features approaches involve further annotating of training data and assigning a word to each section of an image, which could relate to the ABCD rules or dermoscopic features. This can be achieved using a saliency map to separate an image into super pixels and assigning relevant words to each region. This is followed by classifying each area relating to the colour and texture descriptors [17]. However, these answers are often found to be ambiguous and difficult to interpret. This is because each area is assigned a word or multiple words, but do not directly show or enhance feature. Therefore the person using it would need the relevant experience to recognise the features effectively.

Clinically relevant features describe systems that follow a segmentation, features extraction and classification[105] approach. These steps are designed to segment the lesion to quantify the ABCD rules[15, 35] and other dermoscopic features. These hand-crafted features using edge filters can provide a basis for a diagnoses, but, this is sometimes met with scepticism because it is difficult to quantify the features for medical meaning. However, a recent paper has improved interpretability by analysing each feature with an SVM before merging the results using Bayesian fusion. This means each result is visualised and assigned a score similar to TDS possibly improving the diagnoses accuracy. Furthermore, this is a risky area to research because it is beginning to converge with texture extraction and deep learning methods; where clinical features are separated during classification through means of explainability. This means that using hand-crafted features could become irrelevant in the next couple of years.

Texture extraction methods utilise segmentation and border to visualise the ABCD rules; D being Diameter instead of Dermoscopic structure. Smaller dermoscopic features are not made interpretable in these examples and the area is instead extracted as a texture as GLCM, LBP, etc. Therefore all features are analysed at a given time with exceptionally high accuracies. This holds the benefit of having a higher accuracy than extracting clinically relevant features, but holds only the bare minimum of visualising clinically relevant features to doctors. Biases will therefore be harder to recognise possibly leading to rarer lesions being misdiagnosed.

Deep learning models are the current highest accuracy techniques that consist of training models directly from image samples. These methods have been improved to produce a visualisation through heatmaps and loosely visualising some relevant features, but the results cannot be retraced [49], making them un-trustworthy. While there are methods being developed that allow for explainability, the results are still not interpretable[88] enough for use within a clinical environment. Therefore many useful machine learning algorithms should not be utilised within a clinical environment. Furthermore, this style of technique requires a large amount of training data which is not always available, especially for rarer skin conditions. This means there is a likelihood of the results being less accurate

in rarer conditions with no means of proving otherwise.

2.5 Case-Based Reasoning

Case-based reasoning (CBR) is a probelm solving methodology that uses experiences (or cases) to address and solve newer issues, rather than relying on generalized methods. The goal of this method is to use stored prior cases and adapting their solutions to fit new circumstances. CBR is widely used in the medical domain and is mostly used for tutoring new doctors[64]. This process enhcances the learning process.

CBR is essentially a methodology of using previously solved cases to solve new ones. Most of the relevant data tying cases together needs to be sorted manually, as such finding similar cases is not always possible. An automted system that can find relevant features in a skin lesion and find skin lesions diagnosed with similar features would be valuable to doctors.

2.6 Research Methodology

The reason for writing this review was to select the best approaches to skin cancer detection and regarding whether they can be utilised within clinical environments.

Combining the information helps specify what is currently known in literature and highlighting what areas need further work.

The literature includes explainable techniques or have supporting techniques developed for explainability.

2.6.1 Research Questions

The goal of this systematic review is to answer the following questions:

1. What are the major techniques developed for the detection of skin cancer?
2. Are these techniques suitable for use within clinical environments?
3. Can these techniques be supported with other algorithms to improve explainability?

2.7 Neural Network based algorithms

There are widely available studies describing extensive neural network algorithms relating to the detection of skin cancers. ANN, KNN, DNN, CNN, and GAN technqiues were described in the following studies[89, 32].

2.8 Artificial Neural Network (ANN) Techniques

An artificial neural network is a nonlinear statistical prediction technique.

Roffman et al.[82] proposed a multi-parameter artificial neural network (NN) for the early detection of non-melanoma skin cancer (NMSC), specifically in the absence of known ultraviolet radiation (UVR) exposure and family history—significant risk factors. Using data from the 1997–2015 National Health Interview Survey (NHIS) with 2,056 NMSC cases and 460,574 non-cancer cases, the NN was trained with 13 parameters, including gender, age, BMI, diabetic status, smoking status, and others. The study achieved an area under the ROC curve of 0.81 for both training and validation. The results indicated robust performance, with training sensitivity at 88.5% and specificity at 62.2%, and validation sensitivity at 86.2% and specificity at 62.7%. These were similar to other techniques showing that the NN's reliability in early detection with high sensitivity and specificity.

Xie et al.[110] proposed a novel method for classifying melanocytic tumors as benign or malignant through the analysis of digital dermoscopy images. The algorithm uses three key steps: first, lesions are extracted using a self-generating neural network (SGNN), second, features describing tumor color, texture, and border are extracted, and third, lesion objects undergo classification using a neural network ensemble model. The model addresses challenges considering new border features that characterise irregularities on both complete and incomplete lesions. The ensemble classifier, incorporating backpropagation (BP) neural networks and fuzzy neural networks, is designed to enhance overall performance. Experimental validation on diverse dermoscopy databases, including images from different racial groups, demonstrates significant improvements in classification accuracy, showing the effectiveness of the proposed border features and classifier model.

Kumar et al.[51] proposed an enhanced strategy for the early detection of three types of skin cancers and designed for mobile devices. Input for the detection process consists of images depicting skin cancer lesions, categorised as either cancerous or non-cancerous. Image segmentation was found using Fuzzy C-means clustering to segment homogeneous regions, and various filters including Local Binary Pattern (LBP) and GLCM in the pre-processing stage to enhance the quality of the images. For classification, an Artificial Neural Network (ANN) is trained using a differential evolution (DE) algorithm. The proposed ANN-DE method demonstrated superior accuracy and efficacy compared to traditional methods, achieving an overall accuracy of 97.4% with the HAM10000 and PH2 datasets.

2.8.1 Kohonen Self-Organising Neural Network (KNN) Technqiues

Mengistu et al.[60] proposed a self-organizing NN and radial basis function (RBF) to diagnose skin cancers between melanoma, BCC, and SCC. This involved the colour extraction method GLCM, and morphological features from the lesion images. These features were then utilized as input for the classification model. The system outperformed K-nearest neighbor, artifical neural network, and naive-Bayes classifiers with accuricies of 71.23%, 63.01%, and 56.15%,

respectivly. The proposed technique achieved an accuracy of 93.15%.

Lenhardt L, et al.[55] proposed a skin cancer detector based on k-Nearest Neightbors. Various optical specroscopic techniques were extensivly employed over the years as diagnostic tools to differentiate between malignant diseases. The data including gathered from synchronous fluorescent spectroscopy (SFS) and chemometrics were used to train a KNN and an ANN. On the test dataset, KNN exhibited a classification error of 2–3%, while the classification error for ANN ranged from 3% to 4%.

2.9 Deep Neural Network (DNN) Techniques

2.9.1 Convolutional Neural Network (CNN) Techniques

Yunendah Nur Fu'adah, et al.[36] proposed a system for diagnose between skin cancer and benign lesions using Convolutional Neural Network (CNN). The proposed model uses three hidden layers with output channels of 16, 32, and 54 layers. Several optimisers are compared including SGD, RMSprop, Adam, and Nadam. The CNN model employing the Adam optimiser demonstrated the highest accuracy of 99% in classifying the dataset. Using random regulators, the CNN procedure has an accuracy of 97.49%, successfully distinguishing between melanoma, carcinoma, and nevus lesions. Augmentation data from the ISIC dataset played a role in training the model to differentiate between malignant and benign skin cancer lesions. The study's performance outcomes underscore the potential of using the proposed model as a diagnostic tool for medical professionals in identifying skin cancer.

Hasan, et al.[20] proposed an automatic skin cancer detecter using Convolutional Neural Networks (CNN) to classify cancer images into malignant or benign categories. The features of affected skin cells are extracted through the segmentation of dermoscopic images using feature extraction methods, followed by the application of CNN to sort and categorise the extracted features. This approach has an accuracy of 89.5%, with a training accuracy of 93.7% when utilising publicly accessible datasets.

Raja Subramanian, et al.[95] proposed the use of convolutional neural networks (CNN) with an architecture consisting of 17 layers (8 convolutional layers and 5 max-pooling layers and 4 fully connected layers). The study model achieved an accuracy of 83.04% and a precision of 81.86% using the HAM10000 dataset with images the size of 600×450. Various research papers and methodologies were explored. Findings indicated that the standard CNN model has detection accuracy for skin cancer.

2.10 Generative Adveserial Network (GAN) Techniques

Rashid et al.[77] proposed a skin lesion classification technique based on Generative Adversarial Networks (GAN). The system employed GAN generated skin lesion images to

increase the size of the training dataset because current dataset sizes hinder the full potential for medical imaging for classification tasks. The generator module utilised a deconvolutional network, and the discriminator module employed a Convolutional Neural Network (CNN) for classification across seven different skin lesion categories. Comparing results with ResNet-50 and DenseNet showed accuracy rates of 79.2% and 81.5%, respectively. Notably, the proposed GAN-based approach achieved the highest accuracy at 86.1% for skin lesion classification. Results demonstrate that GAN-based augmentation yields perfomance imporvements.

Bisla et al.[20] proposed a deep learning approach incorporating data purification and Generative Adversarial Networks (GAN) for augmentation. This paper aims to generate data increasing the size of the dataset to imrpove data balanced-ness. This technique utilises decoupled deep convolutional GANs for data generation. A pre-trained ResNet-50 model was fine-tuned with the purified and augmented dataset to classify dermoscopic images into melanoma, SK, and nevus categories. The proposed system surpassed the baseline ResNet-50 model, achieving an accuracy of 86.1% in skin lesion classification. This highlights the effectiveness of the integrated GAN-based approach for improving classification accuracy by generating augmented image samples.

Ibrahim et al.[2] proposed a data augmentation technique for skin lesions using Self-attention Progressive Growing GANs (SPGGANs), further enhanced using stablilisation technique. The goal is to increase the size of the data by generating photo realisitc and distinct images of skin lesions. The technique generates fine-grained 256×256 skin lesion images tailed for Convolutional Neural Networks (CNN) detection, overcoming some challenegs of convential GANs. The accuracy The technique achieved an accuraacy of 67.3% and the study suggests the approach can be valuable in clinical practice.

2.11 Explainability Techniques

The techiques mention in this section are ones including deepshap that are used alongside exisiting DNN techniques to make models more explainable.

2.11.1 Interpretable Model-agnostic Explanations (LIME)

Marco Tulio, et al.[**empty citation**] developed an XAI model called Local Interpretable Model-agnostic Explanations (LIME) that approximates any black box machine learning model with a local interpretable model to explain each individual prediction for black-box appraoches. LIME can be applied to any machine learning model, regardless of its architecture or complexity. The primary goal of LIME is to provide interpetable explanations for individual predicitions made by machine learning models. It does this by generating local faithful and simplified models (interpetable surrogate models) around specific instances of interest.

The following is how LIME functions:

1. Select Instance: Choose a specific instance for which you want to explain the model's prediction.
2. Generate Perturbations: Introduce small random perturbations or variations to the chosen instance to create a dataset of slightly modified instances.
3. Get Predictions: Use the complex model to make predictions for each perturbed instance in the generated dataset.
4. Train Surrogate Model: Fit a simple and interpretable model (e.g. linear regression) on the perturbed instances and their corresponding model predictions. This surrogate model serves as an interpretable approximation of the complex model in the local neighborhood of the chosen instance.
5. Interpretation: Analyse the coefficients or rules of the surrogate model to gain insights into the factors influencing the complex model's prediction for the selected instance.

LIME has been utilised in healthcare for the analysis of breast tumour classification[75], and diagnosis of pigmented skin lesions[33]. Its ability to calculate feature importance for machine learning predictions has made it a valuable tool for improving the interpretability of AI models in the medical domain.

2.11.2 Gradient-weighted Class Activation Mapping (Grad-CAM)

Gradient-weighted Class Activation Mapping (Grad-CAM) is a technqiue for visualising the decision-making process of deep neural network (DNN) including convolutional neural networks (CNNs), in image classiciation tasks. This functions by feeding the inpit image through the network where the forward pass is perfoermed until the final convolitional or pooling layer before the fully connected layer. Back propogation captures how much each pixel contributes to the feature maps contribute to the final version. Gradients are predicted with respect to the feature maps of the chosen layer and are then global average-pooled, meaning that the importance of each feature is summarised by taking the average of all its pixel-wise gradients. The weighted sum of the feature map is computed where the weights are determiunted by the global average-pooled gradients. The map is passed through a rectified linear unit (ReLU) activation function to retain only the positive contributions, ignoring the irrelevant or negatively impacting regions. The final heatmap is upsampled to the original input image size which provids a visualisation fo the important regions.

The generated heatmpa hights the regions of the input image that were influential in the model's decision for a specific class. For example in the context of melanoma detection it would be positive to see the focus of the heatmap to be on the lesion itself, where if it was focused on the area of skin it shows the detector isn't necassarly functioning correctly. Grad-CAM is a valuable tool for model interpretation especially in the medical domain where understanding the model is crucial.

2.11.3 DeepSHAP

2.12 Ensemble Learning Techniques

Ensamble learning is a branch of machine learning based on the decision-making process to intergrate it into systems better[111]. Decision-making is the process of making a choice among many options and summarizing evidence to draw a conclusion. An example is case-based reasoning which involves classifying and presenting visually or statistically similar cases and their results.

2.13 Hybrid machine learning techniques

Hybrid machine learning are techniques that aims to use the superior accuracy of deep learning algorithms that are difficult to interpret alongside more explainable algorithms inlcuding bayesian networks, SVMs and others.

2.14 Feature Extraction Techniques

Many CAD frameworks follow a methodology for the classification of skin lesions. These are listed below:

1. Segmentation – Image segmentation is the process of partitioning an image into multiple segments for more accessible analysis. These areas can be separated manually by a dermatologist (known as the ground truth) or separated automatically using statistical or machine learning algorithms.

2. Feature Extraction - Gathering features through filtering, morphology and other statistical approaches. ABCD rules include asymmetry, border, colour, and dermoscopic structures.

3. Combination - Combining the extracted features before using Principal Component Analysis (PCA) or after classification using Bayesian Fusion. Others combine the results using the Total Dermoscopy Score (TDS).

4. Classification – Measuring the results from the features and components through classification. Containing the final diagnosis of the type of skin lesion (Naveus, SK, or Melanoma)

2.14.1 Segmentation

Yading Yuan and Yeh Chi Lo describe a fully convolutional network (FCN) with an accuracy of 91.7% with the PH^2 dataset[112]. FCN is a variation of a CNN using 1x1 convolutions

instead of dense layers. Essentially, an FCN forms a more complex function (generating a more complex neural network), whereas the CNN forms a less complex function, likely to degrade essential features. Therefore, more data is needed to train an FCN effectively than a CNN. After the convolution layers, transposed convolution layers (or deconvolution) and other layers (un-pooling) up-sample the input feature map to the size of the input image. Then, the network, trained from ground truth (human-generated segmentation mask) and the original images, can automatically generate segmentation masks based on textures and colours of the skin lesion provided. There are dozens of examples of this, such as SegNet[14], which is another transposed CNN not designed initially for skin lesions but is effective at segmenting skin lesions.

E. Meskini et al. proposed using Otsu binarisation - a threshold technique that is effective at locating the border of a skin lesion after segmenting using Segnet[61]. Researchers proposed that when analysing the skin lesion border using ABCD rules, the original SegNet methods were ineffective because the ground truth is subjective - ineffective at finding the border cut-off between the skin lesion and skin. While SegNet has a 91.7% with the PH^2 dataset, the data is not effective at finding the precise border cut-off required for accurate border classification using ABCD rules. Therefore, researchers proposed the Otsu threshold to find the skin lesion border after segmenting using SegNet. Fan proposes another technique that uses a saliency-based segmentation approach to capture the area, followed by an Otsu threshold[34] to find the border cut-off from the skin lesion with a precision of 96.78% validated using the PH^2 dataset.

Pedro M.M. Pereira et al. proposed local binary pattern clustering (LBPC) to exaggerate the border, producing accurate results when classifying ABCD rules than ground-truth borders in the PH^2 dataset[71]. Local binary patterns (LBP) are texture descriptors calculated by comparing the centre pixel (of each pixel in the grey scaled image) with the eight neighbouring pixels as 'i', and converting it to a binary using the equation: $[if centroid > neighbour_i = 0, otherwise = 1]$. These eight neighbouring values produce a binary of 01101100 (decimal of 108) and change the centroid to 108. Next, the described process repeats on each other pixel in the image. Finally, the newly filtered image subtracted from the original grey-scaled image creates a segmentation mask with an accurate border cut-off. Finally, Pereira describes classification methods using SVM or FNN presenting the extracted border with an accuracy of 79% and 77% (respectively) with the MED-NODE dataset.

An approach by Albanhli[4] uses a deep learning-based segmentation algorithm using YOLOv4-DarkNet and active contouring for melanoma and skin lesion detection and segmentation. This technique provides a classification of the skin lesion and a segmentation, demonstrating a high level of practicality for clinical decision support systems.

Seeja R D[87] proposed a technique that utilizes a convolutional neural network (CNN) based on a U-net model architecture for the segmentation based on colour, texture, and shapes. The U-net model architecture is a popular choice for image segmentation tasks due to its ability to capture both local and global features effectively.

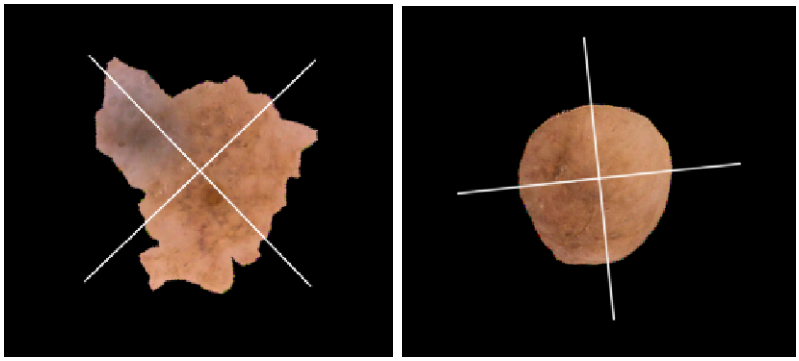


Figure 2.1: Images of two skin lesions from the PH^2 dataset showing the asymmetry calculated from moments.

Hyunju Lee[53] proposed a technique that utilizes an edge fill method called u-otsu for segmentation, using the U channel from the YUV colour space to calculate the histogram. Otsu calculates the optimal threshold value to separate foreground and background pixels based on the histogram of the image.

Another technique by Pedro[71] uses a newly developed technique called Local Binary Patterns Clustering (LBPC). Using a Local Binary Pattern (LBP) filter by subtracting the gray-scale image from the LBP filter after a Gaussian filter, resulting in the creation of a mask. This has been successfully used for the detection of melanoma.

2.14.2 Handcrafted Features

Handcrafted features are the extraction of particular features using statistical algorithms the benefit of separating data into components is a more accessible breakdown, improving explainability. In addition, this might instantiate trust for use within a clinical environment and prove more helpful to doctors.

Asymmetry

Asymmetry can be measured using the bi-fold technique, which involves drawing a line down the middle of the skin lesion and comparing the two halves to confirm whether the sides match, on both the horizontal and vertical axes, as shown in 2.14.2. If the two sides are greatly different, it could be a warning sign of melanoma. Asymmetry can be measured using the shape[113], colour[46], and texture[5].

Measuring the asymmetrical shape requires a precise border cut-off. Ihab S. Zaqout[113] describes a technique using the centroid and rotation of the skin lesion using moments of inertia. By Folding the skin lesion on both vertical and horizontal axes subtracting the opposite half. Pixels that cannot subtract are summed and compared with a threshold

considering the skin lesion asymmetrical if the combined sum is more than the threshold.

Reda Kasmi and Karim Mokrani[46] describe creating a grid of 20x20 pixels of the skin lesion image and converting it into the LAB colour space. Next, each block's average colour is compared with a perpendicular block (vertical and horizontal axes) using the three-dimensional Euclidean luminance distance, a-axis, and b-axis. If more than half of the colour comparisons are over the threshold, that axis is considered colour asymmetrical. Blocks that have no symmetrical pair are ignored. Finally, luminance calculated separately prevents brightness problems. This technique has an accuracy of 94% with a private dataset.

Measuring similarities in texture can be achieved by using SIFT-based similarity and projection profiles[5]. SIFT is scale-invariant and helpful for texture components with varying texture quality. First, the skin lesion is split vertically and horizontally across the centre into four halves, comparing texture components on the symmetrical halves and measuring the similarity. Lastly, the projection profile in the x and y directions generates histograms. These results train a decision tree and have an 80% accuracy of the ISIC 2018 with 204 images privately annotated for ABCD rules and combined.

Ihab S. Zaqout[113] describes a technique using the centroid and rotation of the skin lesion using moments of inertia. By Folding the skin lesion on both vertical and horizontal axes subtracting the opposite half. Pixels that cannot subtract are summed and compared with a threshold considering the skin lesion asymmetrical if the combined sum is more than the threshold.

Kasmi and Mokrani[46] create a grid of 20 by 20 pixels from the skin lesion image and convert it into the LAB colour space. The average colour of each block is compared with a perpendicular block (vertical and horizontal axes) using the three-dimensional Euclidean luminance distance, a-axis, and b-axis. If more than half of the colour comparisons exceed the threshold, they consider that axis to be colour asymmetrical. They ignore blocks that have no symmetrical pair. Finally, they calculate luminance separately to prevent brightness problems. This technique achieves an accuracy of 94% with a private dataset.

Ali[5] uses SIFT-based similarity and projection profiles to measure similarities in texture. SIFT is scale-invariant and helpful for texture components with varying texture quality. First, they split the skin lesion vertically and horizontally across the centre into four halves, compare texture components on the symmetrical halves, and measure similarity. Lastly, they generate histograms for the projection profile in the x and y directions. These results train a decision tree and achieve an 80% accuracy of the ISIC 2018 with 204 images privately annotated for ABCD rules.

Prior studies have introduced techniques that measure distinct aspects of asymmetry, such as Ihab S. Zaqout[113] measurement of shape, Kasmi and Mokrani[46] measurement of colour, and Ali[5] measurement of texture. The new approach seeks to combine the following approaches into a more comprehensive analysis of asymmetry that takes into account multiple features of the skin lesion. The proposed novel technique updates colour measurement to improve accuracy using superpixels and an SVM model.

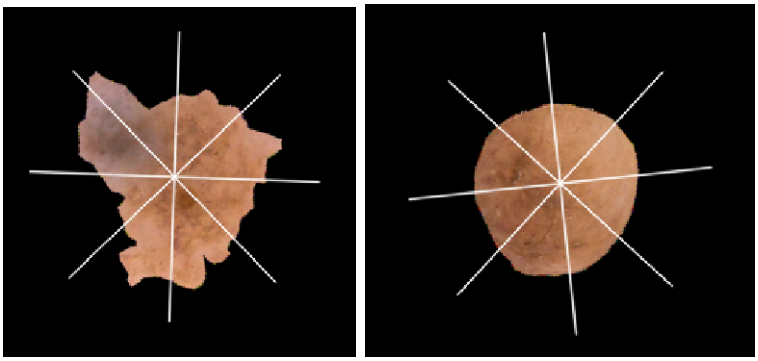


Figure 2.2: Images of two skin lesions split into 8 sections using moments, each border is measured for irregularity.

Border

Estimating border irregularities involves splitting the skin lesion into eight equal sections (through the centroid), where each section with tight corners and convexity is considered irregular. Each irregular section of the border adds a score of 1 ranging from 0 to a total of 8, as shown in figure 2.14.2.

Border irregularity contours were found by splitting the skin lesion into eight segments around the centre, and then calculating a fitting error for each. If the error is larger than 0.05 (x contour), that area is considered irregular[46].

Abder Rahman H. Ali et al. calculate the compactness of each border by first calculating the contour around the area of the lesion containing x and y positions. Next, measure the space between each position to estimate the compactness. The tighter the curves and corners, the more contour positions, revealing irregular borders within a segment, combining all of these scores creates the irregularity index[113].

Fractal dimensions (FDs) are a statistical index measuring the detail in a pattern changing with the image scale index. One technique called box-counting increases values if there are more corners and edges around the border. The higher the value demonstrates the level of border irregularity. Ali describes using machine learning alongside Zernike moments, and convexity measurements for a high-accuracy border irregularity classification[6]. However, results are ambiguous because the output is either “irregular” or “regular” border (not relating to the TDS). Thus, conforming to the TDS and splitting the border into eight sections would make it more interpretable and useful o doctors. However, a hybrid GNB and CNN approach are combined to allow interpretability through GNB.

Colour

Colour refers to the shades of pigment within the area of a skin lesion, not referring to abnormalities relating to bruises, crust, and grazes. Melanoma usually contains more than

two colours compared with benign lesions, singular in colour. Skin lesions can consist of one or many colours: white, red, light brown, dark brown, blue-grey and black.

Finding colour variations has been achieved by calculating the normalised standard deviation of the red, green, and blue components[90]. The normalisation process improves the recognition of normal skin pigmentation, which would show pigmentation levels, making comparisons easier between different skin lesions.

Arthur Tenenhaus, et al. utilise joint learning using Kohonen map, and k-means clustering[99]. Five random pixels create a 5 by 5 Kohonen map represented by 25 neurons in a neural network for each skin lesion in the dataset. Colour variations on a 25-dimensional vector find the proportions of pixels projected onto each of the 25 neurons. Next, K-means classifies the skin lesions set by the number of colours found by dermatologists. Only four colours were present in the dataset in this scenario, while seven could be. Eventually, the colour components are represented as a 42-dimensional vector and are passed into a KL-PLS based classifier to detect variations in colour at 66% using a private dataset.

Reda Kasmi, et al. locate the number of colour variations by converting the image into the LAB colour space matching the colour ranges that can be perceived by human eyes[67], measuring the average colour distribution of the dataset and assigning each colour as a threshold range. Next, the Euclidean distance between each colour threshold is compared with each pixel colour[46], finding the closest matching colour of the six colours. Finally, removing the areas of colour with less than 5% prevents the classification of dots. This approach uses a colour range of white, light brown and dark brown. However, there is a static threshold value for the other colours, which would be unlikely to cover the ranges of the colours, including red, blue-grey, and black.

Dermoscopic structures

Dermoscopic structure refers to structures on the skin lesion, including pigment networks, structureless areas, dots, globules, streaks, white structures, and 22 others (not including sub-types). Variations of pigment networks are more commonly found in melanoma[9] and are therefore a valuable feature for automatic classification. Similarity other features such as milia-like cysts, a sub-type called milia-like cysts (MLC) called cloudy MLC appears more frequently on melanoma than SK, with a specificity of 99.1% specificity[94].

Javier López-Labraca et al.[58] describes a statistical approach to classifying melanoma using dermoscopic structures through Gabor filtering, support vector machines, and Bayesian fusion. This technique uses a form of soft segmentation to find the area of these dermoscopic features. Firstly the structures are located using Gabor filtering using different values to find fissures and globules. Each structure is then compared with a trained SVM model to check the similarity of the detected features. The results from the model are then combined using Bayesian fusion to reach a result of malignant or benign. Finally, training a CNN model alongside an SVM improves the retractability of dermoscopic structures; compared to a standalone CNN model.

2.14.3 Combining ABCD Rules

This section describes combining features from the ABCD rules into a classification between malignant, suspicious or benign after considering all clinical features. Again, meta-data and texture can potentially improve the results.

Maryam Ramezani et al. proposed a method to extract features from ABCD rules storing them in vectors and extracting the texture as a GLCM. First, these 187 features are shrunk to 13 using PCA[76]. Next, the data trains an SVM to classify skin lesions into benign or malignant with an accuracy of 82.2% on macroscopic images using a private dataset.

Other methods output TDS[113, 114], which combines them using: [(Asymmetry x 1.3) + (Border x 0.1) + (Colour x 0.5) + (Diameter x 0.5)]. A statistical model for each ABCD rule outputs a score in the same format. The benefit is interpretability because it follows the diagnostic procedure. The technique achieved an accuracy of 90% using a private dataset.

2.15 Datasets

2.16 Challenges of Explainability

2.17 Conclusion and Future Work

Many techniques utilise ABCD rules to produce an automatic and interpretable diagnosis. Interestingly, many focus on detecting and classifying asymmetry, border, and colour (ABC) or dermoscopic structures, but neither combine the whole ABCD rules into a single framework. Despite dermoscopic structures providing a means of diagnosing problematic forms of melanoma, including mimics (seborrhoeic keratosis)[43], and non-pigmented melanomas. Thus, it would be valuable to combine both into a single system for possibly higher accuracy.

Despite various valuable features, asymmetry rarely utilises techniques other than statistical models. For example, researchers highly focused on border irregularity and dermoscopic structures, leading to hybrid machine-learning models for their assessment. However, asymmetry still utilises statistical approaches to measure and combine shape, colour, and texture. It would be beneficial to transform this data and process it using an SVM, improving accuracy.

Utilising external data, including feeling, touch, age, and location on the body, are helpful to doctors when diagnosing skin conditions, but is not mentioned in any of the discussed techniques. It would be beneficial to implement this data into the decision-making process.

Chapter 3

Dataset Suitability for Melanoma Detection

3.1 Introduction

This chapter contains an analysis of some popular datasets including ISIC 2019, and PH2. The goal is to identify any relationships between skin lesions and patients. Using this analysis 'the dataset' is created using NHS macroscopic images and metadata.

There is an analysis of the distribution of metadata in these datasets, discussing whether they are consistent with the literature. For example, SK is more likely to have milia-like cysts compared to melanoma. So we check if the distribution of data in datasets specifies this. The reasoning for this is that a Bayesian network is used later and trained on datasets based on the distribution of data. Therefore ensuring that the data is scientifically relevant ensures the Bayesian network is functioning correctly. Certain criteria are found to be ineffective and removed because they are not consistent with the literature.

3.2 Other Datasets

The analysis of skin lesions is an especially laboured field, so there are a huge number of relevant datasets to discuss. Public datasets include MEDLINE, PH2, ISIC, and others making a total of 21 open-access datasets containing 106,950 skin lesion images[108]. Out of these datasets, only the PH2 dataset has publicly accessible metadata regarding ABCD rules with a total of 200 images. ISIC is the largest of these datasets, being a combination of many other datasets, with extra annotations from the original.

The datasets listed in table3.2 include the number of images, classes, and metadata. Out of these datasets ISIC 2019, PH2 and 7-Point Criteria appear to be the most promising. PH2 is especially useful because it is the only dataset representing ABCD rules on asymmetry and colour. SKINL2 was considered, but ISIC 2019 was a much larger dataset with more

Name	Year	Image Type	Number	Classes	Metadata
ISIC 2019	2019	Dermoscopic	33,569	8	Age, anatomical site, gender, and diagnosis
PH2	2013	Dermoscopic	200	3	Asymmetry, colour, pigment network, dots/globules, streaks, regression areas, blue-whitish veil
MED-NODE	2015	Macroscopic	170	2	n/a
SD-198	2016	Dermoscopic & Macroscopic	6,584	198	anatomical site, symptoms, duration, morphology, and colour
SKINL2	2019	Macroscopic (unique tool)	376	8	Gender, age, and fototype
7-point Criteria	2018	Dermoscopic & Macroscopic	2000	2	Pigment network, regression, pigmentation, blue-whitish veil vascular structures, streaks, dots/globules

Table 3.1

metadata. SD-198 is publicly available but not accessible.

Overall ISIC 2019 and PH2 are the most suitable datasets. The PH2 dataset is utilised for an analysis of feature extraction techniques such as the detection of ABCD rules and dermoscopic structures. Then the entire technique is analysed using the ISIC 2019 dataset, which is the largest public dataset.

3.3 Issues

One fundamental problem is the overutilisation of private or privately annotated datasets, making a direct comparison of algorithms (especially relating to ABCD rules) difficult. Some are between benign and malignant[61, 46, 6, 5] while others utilise private or never mention any datasets[46, 90, 99, 76, 113]. None compare their ABCD rules, likely because of subjectivity depending on the dermatologists that labelled them. Ideally, more datasets and labels should be public to assess individual rules and reach objective measurements. Until then, testing algorithms conform with malignant, suspicious, or benign. This is espeically true for the PH2 dataset, although it was around before some of these publications it was not used for testing.

Although there is an ISIC 2020 dataset with a total of 44,108 images, its diagnosis is between benign and malignant and other metadata is on atypical melanocytic proliferation, café au lait macule, lentigo NOS, lichenoid keratosis, naevus, seborrhoeic keratosis, solar lentigo, and other/unknown. The metadata is very specific and doesn't match the requirements of the project, so ISIC 2019 is still a better candidate.

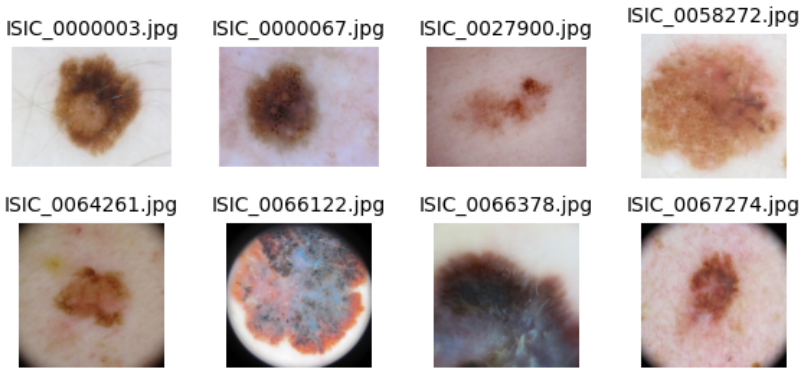


Figure 3.1: Examples from the ISIC 2019 dataset, where first two images are BN, followed by two SK, and four MM.

Name	Year	Total Samples	Differences
HAM10000	2018	11,526	n/a
BCN 20000	2016	19,424	Includes nails and mucosa
MSK	2015, 2017	3918	Coloured stickers covering un-applicable skin lesions

Table 3.2

3.3.1 ISIC

The ISIC dataset is a collaborative effort of many institutions to support the development of automatic classification methods for melanoma detection. The most recent applicable dataset is ISIC 2019, which contains a total of 25,331 images for training and 8,238 for testing, making 33,569 images in total. Each image has corresponding metadata including sex, age, anatomical site, and diagnosis. These images are separated into classes melanoma (MM), melanocytic nevus (MV), basal cell carcinoma (BCC), actinic keratosis (AC), benign keratosis (BC), dermatofibroma (DF), vascular lesions (VL), and squamous cell carcinoma (SCC). This includes segmentation masks.

Images in the dataset shown in figure3.3.1 are captured with a dermoscope and making it substantial for analysing the structures of the skin lesion. However, many of the lesions have incomplete borders, which is especially true for MM because of its increased size to other lesions. It would be a good idea to detect and remove samples that do not have a complete border when analysing ABCD rules.

ISIC 2019 is a combined source of data from different hospital datasets including HAM10000, BCN 20000, and MSK described in more detail in table3.3.1. This is important to mention because each dataset has images captured with differing diagnostic procedures resulting in varying resolutions and styles in which the images are taken. MSK includes coloured stickers covering un-applicable skin lesions that are within the area of the image

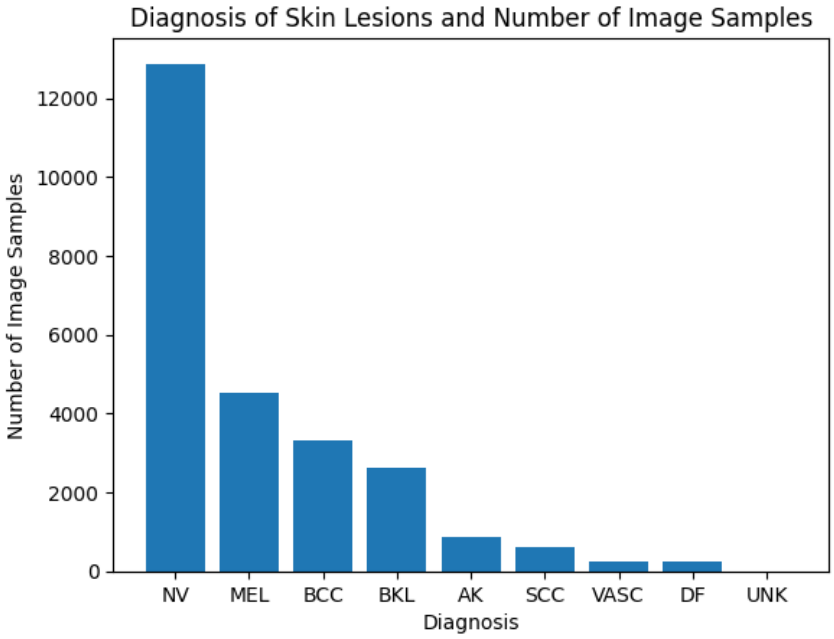


Figure 3.2: ISIC 2019 dataset showing the number of image samples and the diagnosis of those skin lesions dataset appears to be highly unbalanced with half being NV.

and the camera appears to be further away. BCN 20000 contains rarer types of images including nails and mucosa, which do not appear in the other datasets. HAM10000 does not appear to have any notable differences, but it is still captured with different tools and resolutions.

As demonstrated in figure3.3.1 the dataset is highly in-balanced based on the diagnosis of the skin lesion with 12,875 NV and 4,522 MEL. There are only 867 AK, where AK has a very high variance compared to other skin lesions and the sample size is unlikely for adequate detection. Seborrhoeic keratosis (SK) is not in this dataset and AK a similar lesion is described instead. There are only a handful of images for DF, and VASC. The difference in image samples makes the dataset primarily useful for testing between NV and MEL.

Metdata

In this dataset, the images are accompanied by metadata describing patient information. This includes patients age, sex, and anatomical location. Analysing this data provides further insight into the influence of patient information on the classification process. Most of all we are looking for similar numbers for each category and some cases might be removed to balance the dataset and improve classification results.

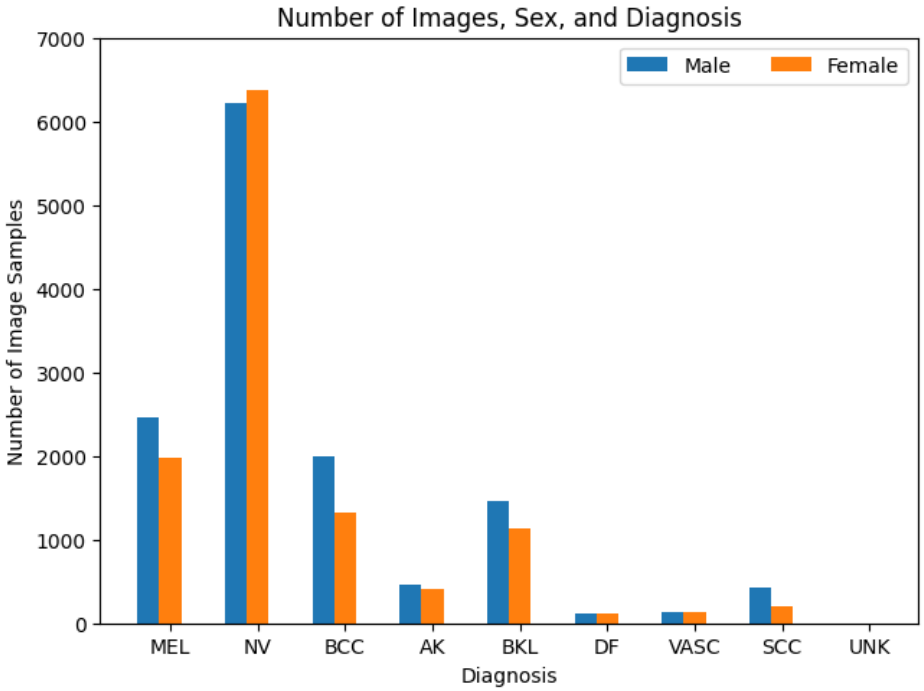


Figure 3.3: The dataset has more male than female patients except for NV which has more samples.

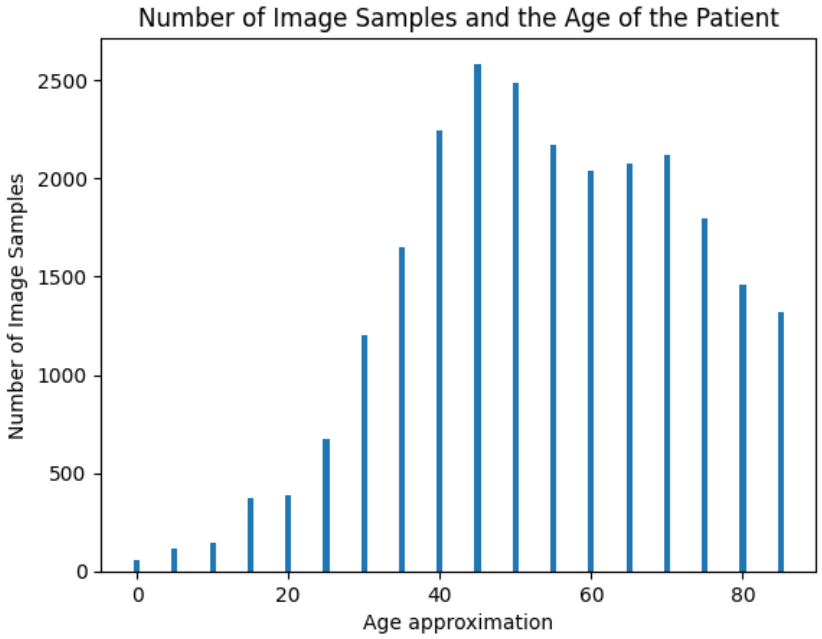


Figure 3.4: This shows the number of image samples compared to the age, the dataset is largely unbalanced regarding age where patients are between 40 and 75 years of age.

Figure3.3.1 demonstrates the number of samples relating to the patient's sex. There are more samples for each type of skin lesion except for NV which there are slightly more female samples. They are all within a very close range and are unlikely to need rebalancing.

As demonstrated in figure3.3.1 the dataset is unbalanced relating to age and type of skin lesion. Variation is likely a result of skin lesions being more likely to develop in older people than younger ones. This might mean that many of the skin lesions are developed and there is going to be unlikely to find underdeveloped skin lesion samples.

In figure3.3.1 the age approximation (in intervals of 5) was compared with the diagnosis. The black line (whisker) represents the minimum and maximum range of age, the box (quartile) shows the interquartile range (IQR), and the centre line in the middle represents the median. Some dots represent outliers in the data, that are outside the age range.

Each class in the diagram is a diagnosis associated with the age of each patient. Interestingly, represented in this data SCC, BCC, and AK appear to develop more in older adults with a median of age 70. Many younger patients were diagnosed with NV with a median age of 46. Whilst MM appears to be diagnosed in adults with a median age of 60. This is correct when regarding literature[**empty citation**].

The comparison between diagnosis and anatomical location provides further insight into the variety of samples. Figure3.3.1 demonstrates the percentage of image samples (based on

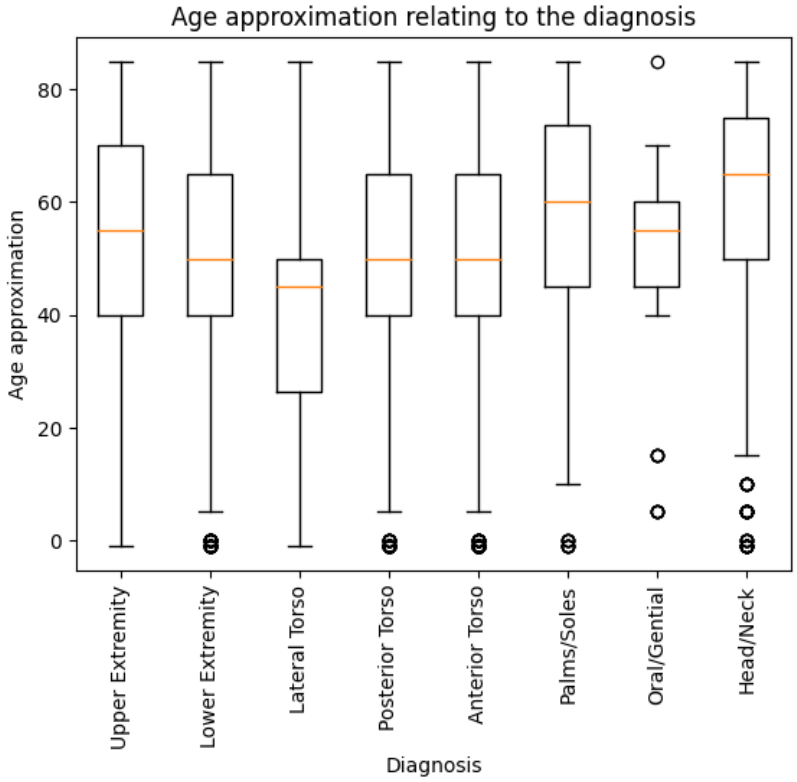


Figure 3.5: The approximate age range of patients and their diagnosis.

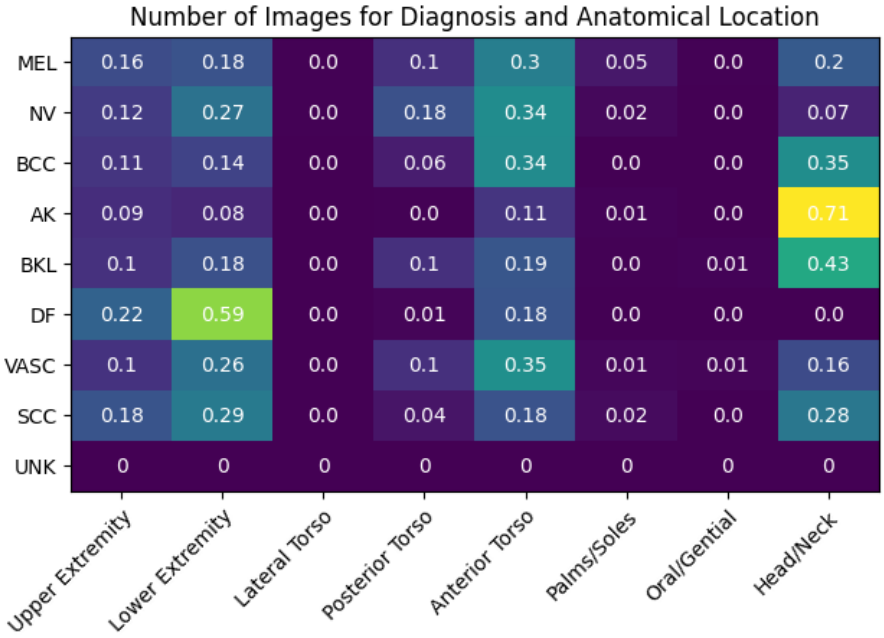


Figure 3.6: Shows image examples associated with the anatomical location and age of the patients.

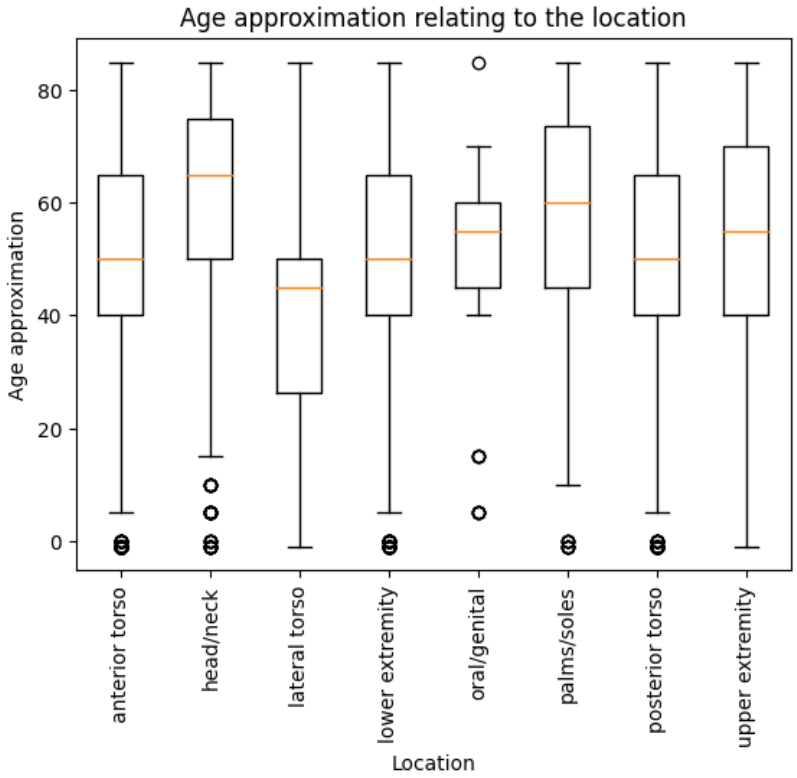


Figure 3.7: Shows image examples associated with the anatomical location and age of the patients.

the diagnosis) there are for diagnosis and anatomical location. Interestingly, AK appears on 69% of images on the head/neck. So there is a strong likelihood that skin lesions are overlapping facial features. Furthermore, DF appears more on the lower extremity and upper extremity at 58% and 22% respectively. Both AK and DF are consistent with the literature[**empty citation**].

Another interesting finding is that most skin lesions are in areas of the body that are frequently exposed to the sun, being anterior torso, head/neck, and lower extremities.

Figure3.3.1 is similar to3.3.1, except it compares the approximate age and location of the skin lesion. There are more older patients who have been diagnosed with skin lesions on their head/neck and younger for the lateral torso. It is concerning that there is a distinct lack of younger patients for palms/soles and head/neck, which could mean more developed skin lesions in these criteria.

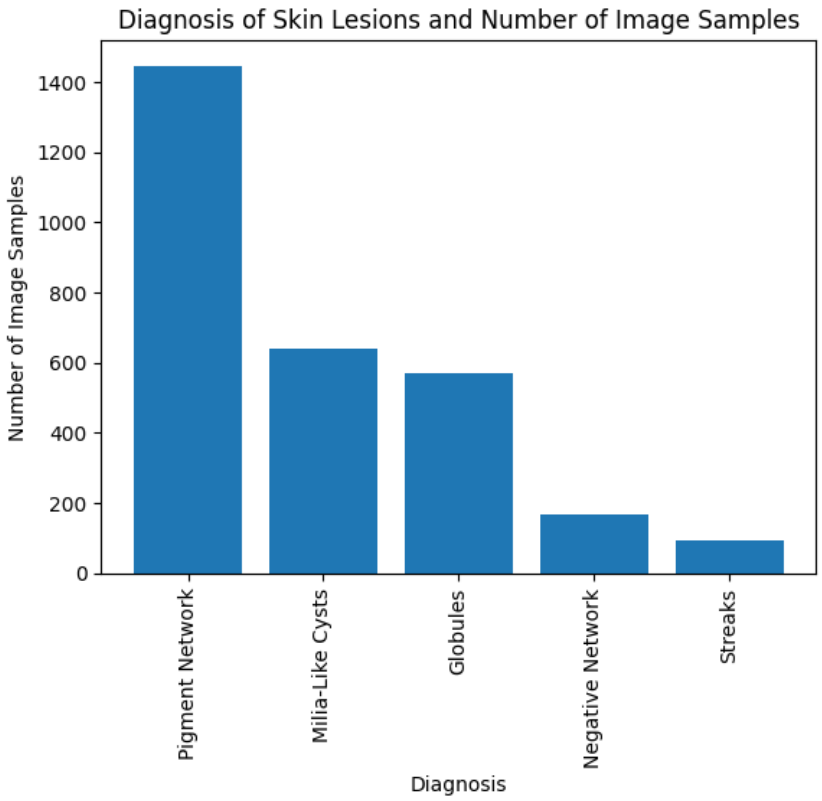


Figure 3.8: Number of images containing dermoscopic structures.

Dermoscopic Structures

ISIC 2017 shares some images with ISIC 2018 including some additional metadata relating to dermoscopic structure. This includes 2,694 segmentation masks of pigment networks, negative networks, globules, milia-like cysts, and streaks. While the original in ISIC 2017 only has metadata for dermoscopic structures, it was linked to ISIC 2018 using image file names to get their diagnosis. The diagnosis is only between benign naevi, seborrheic keratosis, and melanoma.

The dermoscopic structures described in figure3.3.1 show the number of image samples for each dermoscopic structure. Naturally pigmented networks have more than 1400 images which makes it ideal for training a SegNet algorithm. Other demroscopic structures are lacking, such as streaks has less than 100 images. This is too small for most machine learning algorithms.

As described in figure3.3.1 certain dermoscopic structures are split almost evenly between melanoma and benign naevi, except for streaks and negative networks being more common.

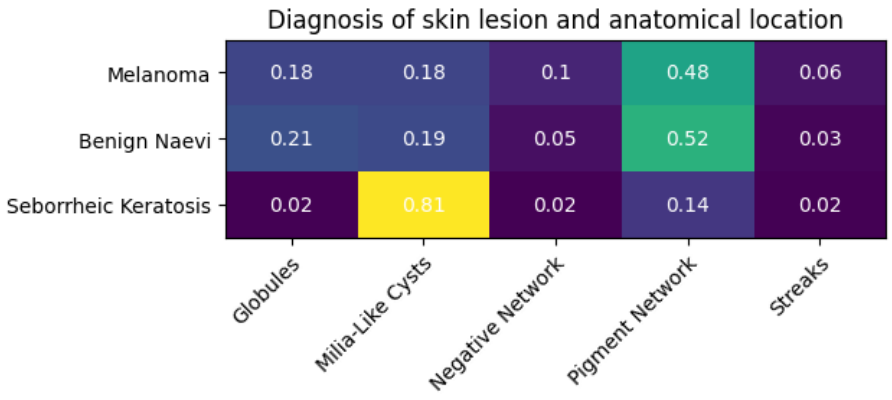


Figure 3.9: Number of dermoscopic structures relating

This demonstrates the importance of being able to detect the difference between typical and atypical pigmented networks. Furthermore, milia-like cysts appear more frequently in seborrhoeic keratosis and there is a lack of pigmented networks. These are again consistent with literature.

3.3.2 Image Assessment

The ISIC 2018 dataset is the latest out of the ISIC datasets to have segmentation masks. One of the issues with the segmentation masks are the use of expert and approximate borders that are mixed.

Summary

In summary, the ISIC dataset including data from 2017, 2018, and 2019 makes this dataset the largest public dataset for skin lesion analysis and melanoma detection. It contains a large collection of dermoscopic images with 8 different diagnoses. The dataset having 33,569 makes it ideal for a diverse range of research and development purposes including the evaluation of machine learning and deep learning models. It also contains 2694 images of dermoscopic structures labelled in the ISIC 2017 version of the dataset. Overall, this is the best dataset currently publicly available for the analysis of skin lesions.

Furthermore, data distribution comparing diagnosis to the age, anatomical location, and dermoscopic structures appears to be consistent with literature[**empty citation**], likely due to the size of the dataset. Otherwise, some image samples have incomplete borders. So, some images should be detected and removed to properly utilise asymmetry, border, and colour.

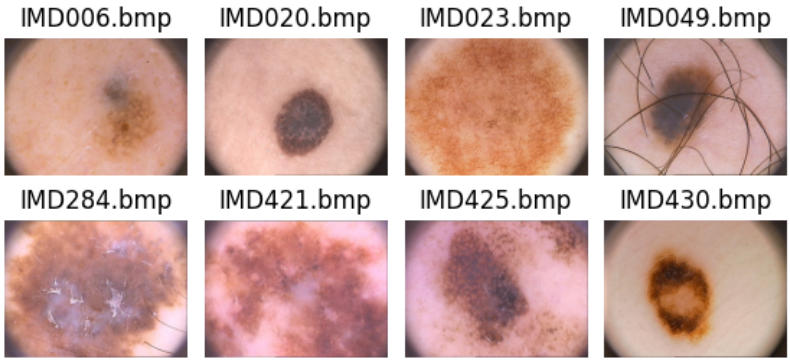


Figure 3.10: Example of images from the PH2 dataset. The first two are standard, the second two are atypical, and the last 4 are melanoma.

3.3.3 PH2

The PH2 dataset is a collection of dermoscopic images that were made available in 2013 by Mendonca, et al[**empty citation**]. It consists of 200 images including 80 common nevus, 80 atypical nevus, and 40 Melanoma. Although the dataset is small it holds substantial metadata for describing features within the skin lesion, including asymmetry, colour, pigment network, dots/globules, streaks, regression areas, and blue-whitish veil. This is the only dataset that has such substantial data regarding the mentioned features. Each image has a segmentation mask of the skin lesion.

The images in figureph2-example-images are some examples of skin lesions from the PH2 dataset. All the images have a circular border and melanoma samples are too big to fit inside the area of the dermoscope in many cases. This results in an incomplete border making it difficult to analyse asymmetry and border from ABCD rules.

As shown in figure3.3.3 there are 200 image samples in total with 80 common nevus, 80 atypical and 40 melanoma. The number of image samples is too small for most neural network techniques. The dataset is highly unbalanced with only 40 melanoma images and 160 naevus images. With such a skewed distribution of classes, the model might become overly biased towards the majority class (naevus), leading to poor performance in identifying melanoma. Another issue is the size of these classes, the scarcity of samples will likely result in overfitting when training models. For this reason, this is not a reliable candidate for training machine learning algorithms when considering that more diverse candidates such as ISIC exist.

Metadata

The benefit of this dataset is the rich metadata allowing for the analysis of specific features within an image and the relationship between them. This allows for the development

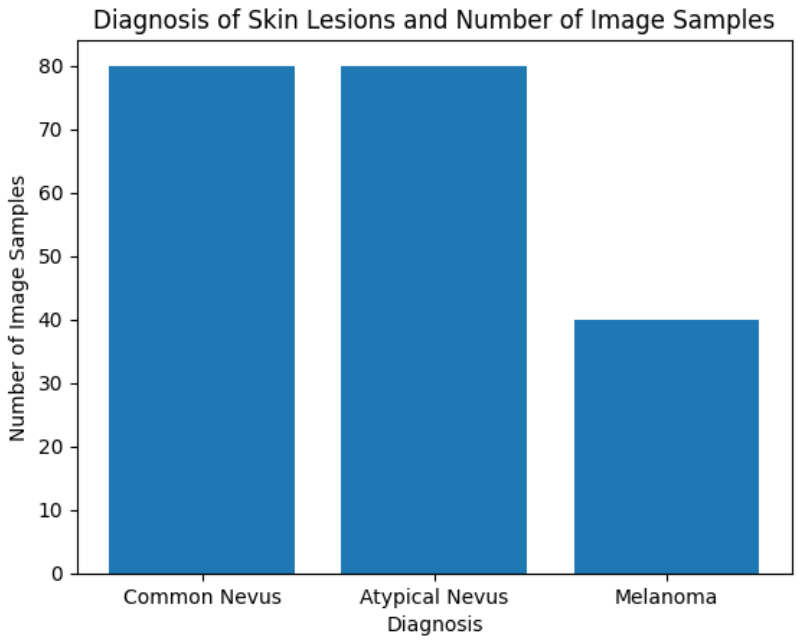


Figure 3.11: Number of image samples and diagnosis in the PH2 dataset.

of more sophisticated algorithms that provide further insight into the characteristics of melanoma and naevus.

Demonstrated in figure3.3.3 demonstrates there is substantial data for measuring the asymmetry score based on the total dermoscopy score (TDS). Typically the algorithms used to measure asymmetry such as bi-fold do not require any training, making the smaller sample size ideal. However, there is a very small sample size for both TDS of 1 and 2, and it would be beneficial to have more.

The observation in figure3.3.3 shows the percentage of images associated with each colour. Light and dark brown are commonly associated with typical naevus. On the other hand, white, blue, and black are more common in melanoma that indicate structural and vascular irregularities. This is true to literature where melanoma is more likely to have a range of colours. The scarcity of red samples demonstrates that it is uncommon in nevus and melanoma. Red is certainly more common in melanoma, but with only a sample size of only 9 the data is likely too stunted to demonstrate this. Other lesions including BCC are more likely to contain red[**empty citation**], but these lesions are not included in this dataset.

There are many records of pigment networks, and dots/globules, but as seen in Figure3.3.3 there are roughly 20 samples for each streaks, regression, and blue-whitish veil. The data is highly unbalanced, so it will be difficult to train a machine-learning algorithm

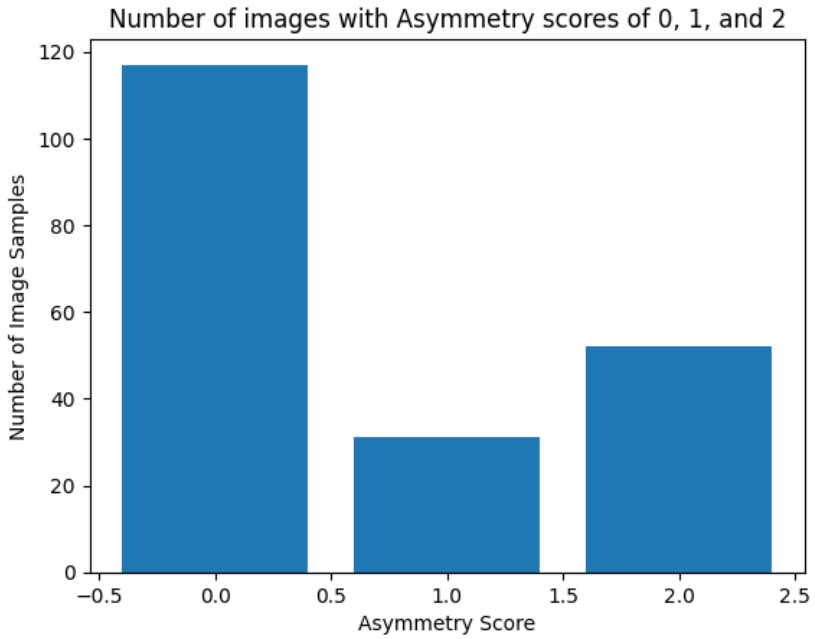


Figure 3.12: This shows the number of image samples and asymmetry score based on Total dermoscopy score (TDS).

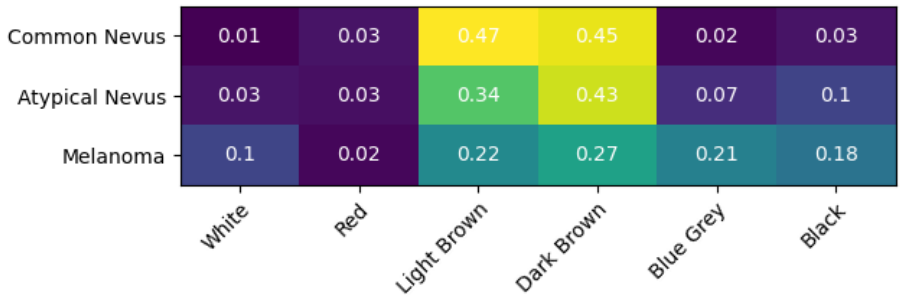


Figure 3.13: Number of colours in the PH2 dataset compared with the diagnosis. Colours are in order white, red, light brown, dark brown, blue-gray, and black.

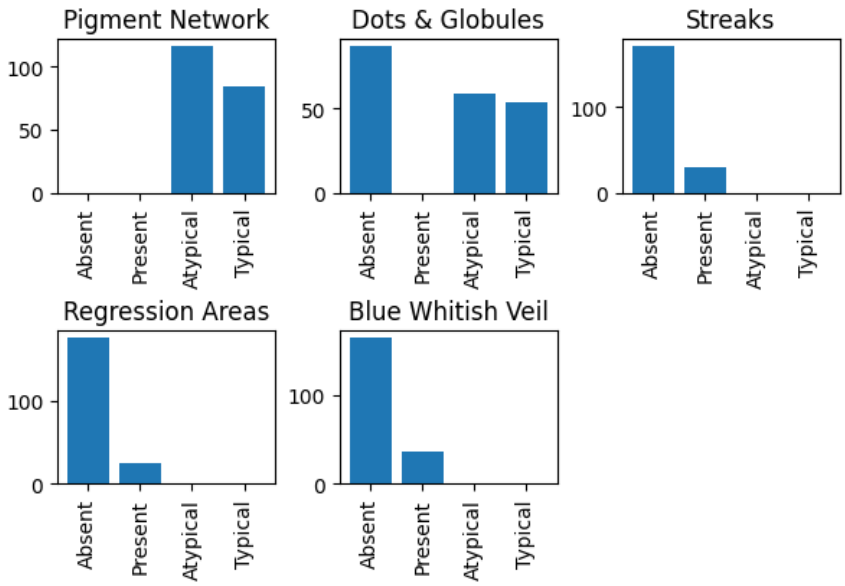


Figure 3.14: Dermosocpic structures and the number of images. These are labelled between absent, atypical, present, and Typical.

for these features. There are more samples of pigment networks, and dots/globules because common. For this reason, typical and atypical features are a good indication of whether the skin lesion is melanoma.

Figure3.3.3 shows dermoscopic structure labels relating to the diagnosis of the skin lesions. Common nevus have typical and present dermoscopic structures, atypical nevus have just as many absent with more present and atypical. Melanoma has more present and atypical types of skin lesions.

Figure3.3.3 demonstrates that pigment networks are present in both nevus and melanoma. Furthermore, streaks, regression areas, and blue-whitish veils are more common in melanoma. Pigment networks and dots/globules use different labels of typical and atypical so it is understandable they do not change between lesion types.

There is little to no overlap between typical and atypical pigment networks between melanoma and naevus. Shown in figure3.3.3 common naevus are all typical pigment networks, while atypical naevus and melanoma are labelled atypical. This demonstrates that testing for pigment networks should be on type (typical and atypical) instead of whether they are present. Unusually, there is very little overlapping in the data. This is very unusual unless it was designed this way purposely, but it would have been more useful to have more samples without pigment networks.

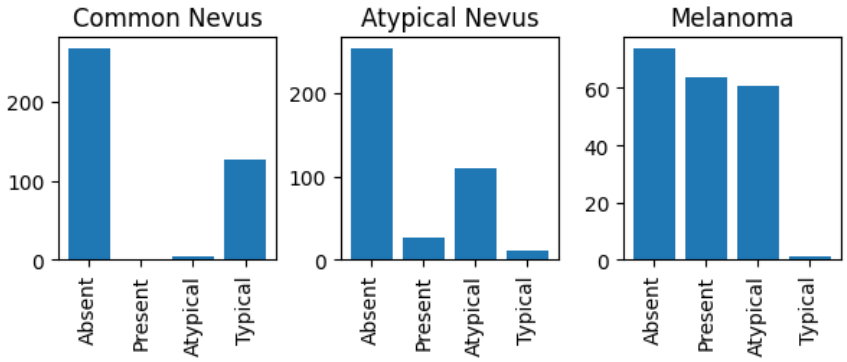


Figure 3.15: Shows the labels of dermoscopic structures, number of images, and diagnosis. These are labelled between absent, atypical, present, and Typical.

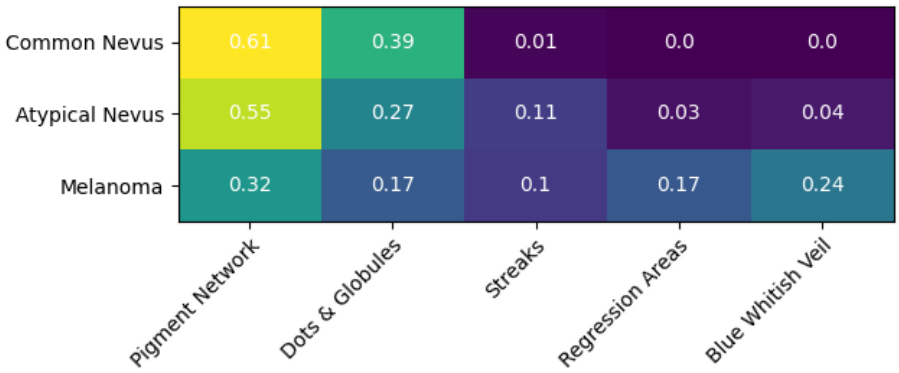


Figure 3.16: Shows the number of images based on the diagnosis and dermoscopic structures present, typical, and atypical.

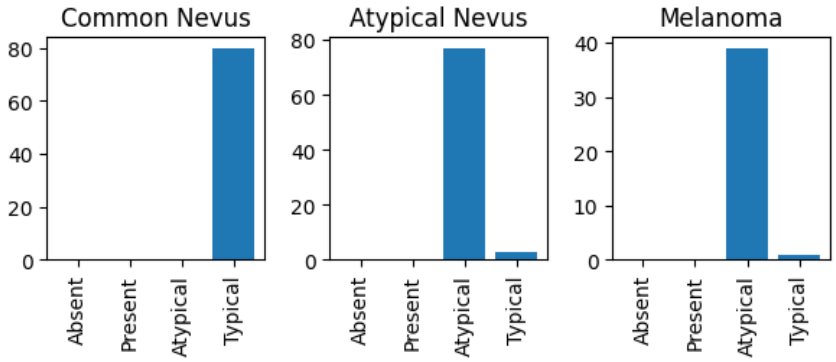


Figure 3.17: Pigment network data relating to the diagnosis.

Figure 3.18

3.3.4 Diagnosis and Image Assessment

Summary

In summary, the PH2 dataset is a valuable resource for researchers considering it is the only one of its kind to provide metadata relating to asymmetry, colour, and dermoscopic structures. This has been used frequently in various studies to develop and evaluate algorithms for skin lesion analysis. Such datasets with substantial metadata are useful for producing explainable results. As explainable (XAI) becomes more common datasets describing clinical features will be necessary. However, one of the downfalls of this dataset is the unusual labelling for dermoscopic structures where the pigment network and globules are labelled between typical and atypical and others are between present and absent. Some metadata was inconsistent with literature including the colour red, which is more common for melanoma, but the data does not demonstrate this. Furthermore, almost every image sample has a pigment network (although common), it would be more useful to have some without this feature.

An issue with this dataset is that many of the papers that utilise ABCD rules and dermoscopic structures do not test their algorithms against this dataset, regardless of the dataset being around before the time of publication[46, 90, 99, 76, 113]. Instead they tend to use privatly annotated datasets, which makes it difficult to replicate results. After analysing PH2 dataset it is likely the small size, lack of red colour samples and the unusual labelling of dermoscopic structures are the reason forthis dataset being avoided. Furthermore, the ABCD rules tend to be subjective and are only relevant to the institution which they are used, therefore this is likely again why the technique is being avoided and why many papers tend to annotate their data. Hopefully some time in the future objective measurements can be accepted and datasets including vast data like PH2 will be accepted.

3.4 Conclusion

Overall both the PH2 and ISIC 2019 provide adequate data for the testing and developing algorithms. ISIC 2019 is suitable for training deep learning algorithms with a total sample size of 33,569 images with relevant metdata. PH2 is certaintly the weakest link with data samples of only 200 images between benign naevi and moles and many papers refuse to test using the dataset and resort to privately annotating their own data. Regardless PH2 appears to be suitable for analysis apart from the lack of red samples.

Considering the data distribution of the dataset both samples appear to be consistent with literature. For example in ISIC 2017 dermoscopic structures, SK has more samples with milia-like cysts and less pigment networks. Furthermore melanoma and benign naevi

have roughly equal pigment networks and less milia-like cysts. Another example is the colour distribution in PH2 while brown and light brown are common in benign neavi, MM has less samples of light and dark brown with more colours of blue, black, and white. The only sample that appeared to be wrong was the colour red, where there was a severe lack of samples.

3.5 Developing the NHS dataset

Whilst recognising the benefits of the ISIC 2019 and PH2 datasets we can begin to develop 'the dataset' using NHS data. The inclusion of NHS data brings real-world clinical cases into the mix, and techniques developed during this project can be tested in the scenario it is intended to be utilised.

Requirements are decided below to highlight potential biases and issues, and then 2,000 images are chosen to create the dataset. This is followed by an analysis similar to the previous datasets and interesting findings from the data.

3.5.1 Requirements

The use of machine learning algorithms for the detection of melanoma is a promising and evolving field with detection accuracies often beating that of dermotologists[10]. However, the effectiveness of such depends heavily on the quality of the datasets used to develop them[96]. The goal of this section is to describe and document the data extraction process from the National Health Service (NHS) and highlight biases, pre-processing, and other potential issues involved in the training of machine learning algorithms for the detection of melanoma. Requirements are first highlighted before gathering the data and are listed below.

There are requirements for this project, which include the use of macroscopic images instead of dermoscopic images. Macroscopic is described as viewing with the naked eye or by taking a picture with standard lenses. When referring to Dermoscopic images, are images captured with a specialised tool called a dermoscope that removes lighting variegation and improves the visual features within the skin lesion usually called dermoscopic features.

Although macroscopic images are used it is important to note that dermoscopy improves the diagnostic accuracy of dermatologists for melanoma when compared with macroscopic examination[109] and is widely considered superior[100]. Dermoscopic images provide a detailed visualization of patterns and structures on the surface of the skin lesion that might not be visible to the naked eye[100]. Some of these structures are pigment networks, asymmetry, irregular borders, and other features that support the differentiation between benign and malignant lesions[100].

Another example shows the diagnosis for BCC was 91% when using dermoscopy, compared to 57% when using close-up images[28]. Similarly, the sensitivity for SCC was 77%

with dermoscopy, compared to 70% with close-up images[28]. These findings highlight the superior diagnostic performance of dermoscopy compared to macroscopic.

The dermoscopic examination is superior to macroscopic examination, however, the project use case specifies macroscopic. The logic behind this is that the tool is specifically designed for general practitioners who are unlikely to recognize dermoscopic features, so there is no need to supply them with dermoscopes. This appears to be consistent with an author's findings showing that 92% of dermatologists correctly recognize at least four size types of melanoma. In contrast, only 38% of non-dermatologists were able to recognize the same number of melanomas[96]. Therefore, 'the dataset' is created with macroscopic images for examination.

Considering the use of macroscopic images there needs to be a more thorough clean-up of the data for it to be used effectively. This will include removing hair and specular reflections to improve classification accuracy. This chapter will discuss the data transformation of NHS macroscopic images, including augmentation techniques to remove lighting, hair and other anomalous data from the images. All of which will support improving the accuracy when classifying.

3.5.2 Data Biases

The use of datasets is fundamental to the development and evaluation of machine learning algorithms, and the accuracy and effectiveness heavily weigh on the quality of the data used. Biases can arise from data collection procedures and pre-processing techniques. Not considering possible biases greatly affects machine learning algorithms using them and their effectiveness. Furthermore, careful consideration is essential to ensure the accuracy and reliability of the conclusions proposed in this document. Failure to consider all these factors could result in skewered conclusions that could undermine the validity of findings. For these reasons, it is essential to carefully identify and evaluate data before using and testing it.

NHS datasets contain a wealth of information that can be utilised. However, some biases need consideration before creating a dataset. These biases include:

1. The diagnostic procedure dismisses skin lesions without recognizably suspicious features and does not reach the phase that photographs were captured. As such, there is a lack of typical benign skin lesions within the dataset, and most have some undesirable features.

2. Dermatologists and general practitioners have diagnosed the large majority of skin lesions which have varying accuracy depending on their experience. Images include metadata on the department and person capturing the image, so the doctors' experience can be measured.

3. Dermatologists could diagnose during an in-person examination where patients can be asked questions in real-time and further tests can be made involving touch. Otherwise,

dermatologists diagnose using previously saved images, which might be less accurate because they lack the insight that an in-person examination would provide.

4. Some skin lesions within the dataset lack metadata including their diagnosis. Such image samples should be avoided.

5. Diagnoses of skin lesions are written in plain text including question marks where there is some uncertainty and the possibility of multiple diagnoses. Only diagnoses that are certain of their findings are used.

6. Photographs of the skin lesions may be captured on different body parts such as hands, legs, face, and others. Most pre-processing methods are designed to differentiate between skin and skin lesions, so it is important to avoid using these images. Otherwise, new pre-processing methods will have to be made and tested.

7. Seborrhoeic keratosis (SK) has similar features to that seen in malignant skin lesions. Therefore, there might be skin lesions diagnosed as melanoma that are SK. Furthermore, because of its similarity, there are many SK images. It will be vital to separate these.

Following the potential issues 2,500 images were chosen from the NHS database. Eight types of skin lesions were chosen Malignant Melanoma (MM), Seborrhoeic keratosis (SK), Atypical Naevi (AN), Benign Naevi (BN), Squamous Cell Carcinoma (SCC), and Basal Cell Carcinoma (BCC).

1. Malignant Melanoma (MM): A type of skin cancer that arises from melanocytes, which are responsible for the pigment melanin (brown skin).

2. Seborrhoeic Keratosis (SK): A non-cancerous growth that originates from cells called keratinocytes.

3. Atypical Naevi (AN): This refers to an unusual or atypical mole that shares characteristics of skin cancer, but they are not cancerous themselves

4. Benign Naevi (BN): Normal benign mole that most people have.

5. Squamous Cell Carcinoma (TM): A form of skin cancer that develops from squamous cells in the outer layer of skin.

6. Basal Cell Carcinoma (BCC): The most common type of skin cancer, that develops from basal cells located in the outer layer of skin.

The database system where the skin lesion images were located uses fotoware software. While there are a substantial number of images roughly reaching 20,000, these were obtained using diverse methods including dermoscopic and macroscopic. Other issues included and were removed:

1. Dermoscopic (Not used in this study)
2. Duplicates (keeping one)
3. Skin lesions were not visible
4. Abnormalities, edge of an ear or belly button
5. Angled or far away images
6. More than 2 skin lesions
7. Almost entirely covered with hair
8. Tattoo
9. Scars
10. Incision

Some images were angled and far away which made it hard to get a clear view of the skin lesion. Others contained multiple skin lesions making it difficult to differentiate which one was being diagnosed. Others including tattoos and incisions were considered more extreme cases and deliberately excluded because there were not enough samples or outside the criteria of the project, respectively.

To remove the mentioned samples a search criteria was used, the example below is for finding melanoma:

("01 Close-up " -"Dermoscopy" -"eye lid" -ear -Nose -scalp -lip -cheek - scar -toe) AND
(-SCC -BCC -"Seb k" melanoma -"atypical mole" -mole)

Some samples could not be removed automatically so they were removed manually by looking through the images. After finding all the samples, the following

1. BN = 500 (600)
2. AN = 500 (600)
3. MM = 400 (500)
4. SK = 200 (264)
5. BCC = 200 (300)
6. SCC = 200 (203)

Attribute	Description
Image ID	an integer representing the image ID - example: 123456.xxx
Doctor	a string representing the forename and surname of the doctor that made the diagnosis - example: JOHN SMITH
Department	a string representing the name of the department - example: DERMATOLOGY
Studio	a string representing the name of the studio used to capture an image of the skin lesion
Capture date	an integer representing the date the image was captured - example: 00:00:0000
Hospital ID	a string representing the hospital ID
Gender	a string representing patient gender
Date of Birth	an integer representing patient date of birth - example: 00-00-0000
Surname	a string representing patient surname
Forename	a string representing patient forename
Initials	a string representing patient initials
Patient ID	an integer representing patient ID
Subject (Tags)	an array representing method the image capturing method (Dermo or Close-Up) and anatomical location
Creator	a string representing the forename and surname of the photographer - example: JOHN SMITH

Table 3.3: This table shows the metadata in each image and a description of each label. Rows highlighted in red are removed to protect patient confidentiality.

A significant difference between this and other datasets is the inclusion of benign naevi and atypical naevi. An atypical mole is an unusual naevi that has features similar to cancer but is non-cancerous. This will provide further insight into their distinguishing characteristics and into the difficulty of diagnosing atypical naevi from other skin lesions.

In the next section, further analysis of the images using the metadata will be conducted to remove image samples with uncertain diagnoses and to balance the dataset for better use with machine learning algorithms.

3.6 Data Transformation and Analysis

Each image holds metadata shown in table 3.6 as EXIF Tags within each image. After extracting the images the metadata included in the images are Filename, Tags (Capturing method, anatomical location), Gender, DOB, Department, Consent, Historical Diagnosis, Diagnosis, and Date Photographed. There was other metadata with each image, but they are potentially identifiable, so to protect patient confidentiality, they were removed.

This is a generalised diagnosis, alongside these images are historical diagnoses of the skin lesion, written in plain text in some cases this question marks when the doctor is uncertain of a diagnosis and other times it includes a slash and a different diagnosis. Although it is not a specific format making it difficult to process, it includes a wealth of knowledge.

Image ID	Historical Diagnosis
998444.jpg	SEB K
549982.JPG	AK / SCC
824466.jpg	ATYPICAL MOLE / ? MM
879067.jpg	? ATYPICAL NAEVI
1028628.jpg	? MM / ? BCC
154414.jpg	1) ? SCC 2) SBCC 3) ? SPOT
739199.JPG	(1) BOWEN'S DISEASE (2) SUPERFICIAL BCC
586010.JPG	SUSPECTED N. MM

Table 3.4: Examples of historical diagnosis and doctors and some unique variations of labelling.

3.6.1 Historical Diagnosis of Skin lesions

The historical diagnosis is written text by the doctor that includes the possible diagnosis. The format is dependent on the doctor, so it changes dramatically. Generally, this is a '?' to show uncertainty and a '/' followed by another diagnosis. All of the variations in the format are shown in table3.6.1.

Some historical diagnoses contain a '?' showing uncertainty from the doctor. Figure3.6.1 describes the number of skin lesions between uncertain and certain. Interestingly AN, BN, and BCC appear to be certain with 117, 36, and 39 respectively. Followed by SK and SCC are roughly half of the images at 106 and 87. Most of all MM shows that more than half of the diagnoses at 259 are uncertain out of 184 that are certain. This in turn demonstrates the type of skin lesions that doctors are having difficulty diagnosing where melanoma is especially difficult.

Multiple diagnoses are sometimes shown, figure3.6.1 shows the number of skin lesions with multiple diagnoses mentioned in the historical diagnoses. Interestingly the most commonly associated are AN with MM at 44 and SK with MM at 22 images. Others are SK with MM, Bowen's disease, lentigo, warts, and SCC demonstrating that SK is associated with the widest range of skin lesions and the difficulty diagnosing it.

Considering that SK has multiple diagnoses mostly for MM and SK, it is a good idea to compare these images and see whether there are any distinguishing features. Demonstrated in figure3.6.1 SK border and colours change dramatically between different lesions demonstrating how difficult it is telling them apart[**empty citation**]. The ABCD rules and TDS is assigned to each image to see whether this diagnostic procedure will be suitable for separating these skin lesions.

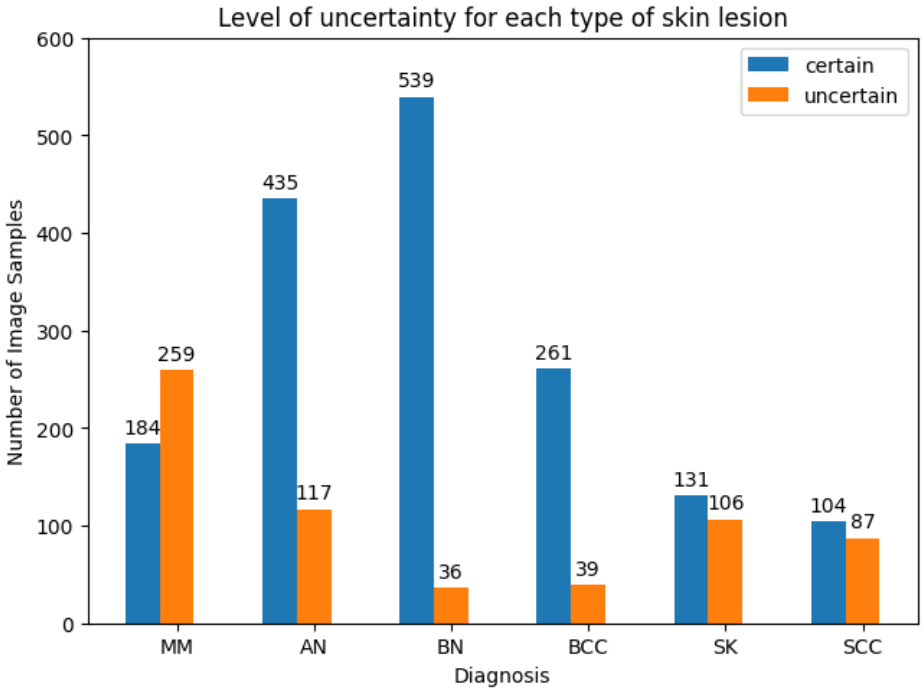


Figure 3.19: Number of image samples relating to the historical diagnosis. Labelled as uncertain if there is a '?' in the diagnosis.

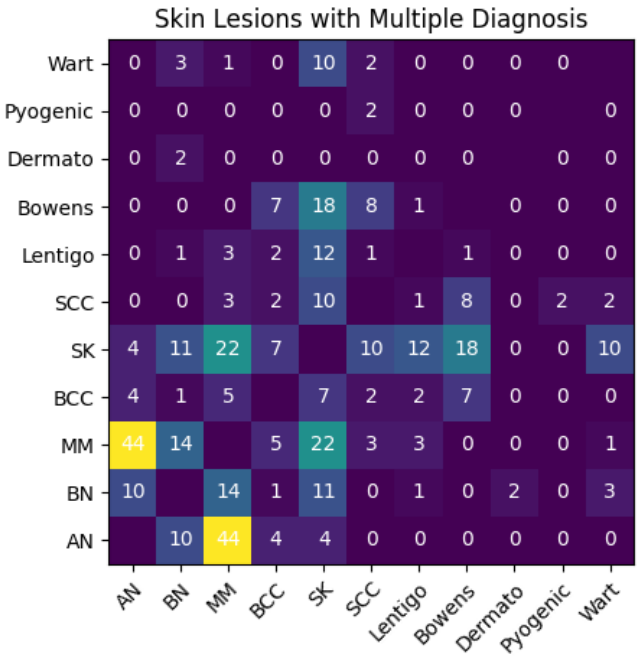


Figure 3.20: Number of skin lesion samples with multiple diagnoses in the historical diagnoses. Other types including lentigo, Bowen's disease, dermatofibroma, pyogenic granuloma, and wart are only associated with the main diagnoses (AN, BN, MM, SCC, BCC) because they are not specifically searched for. This means they are only found in association with the mentioned main diagnoses and this data is likely missing data comparing the other types.

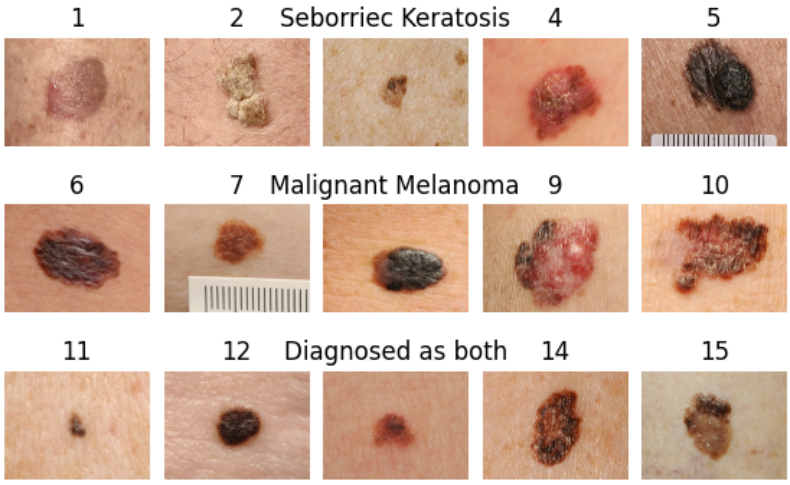


Figure 3.21: Comparing skin lesions that are diagnosed as MM, SK, and considered both MM and SK.

3.6.2 Anatomical location

Anatomical location has a total of 28 different descriptors for specific body parts including leg, wrist, neck, etc. To make comparisons easier with the ISIC dataset each label has been assigned to specific areas such as upper extremity, anterior torso, etc. All the locations are listed in the table3.6.2.

Category	Organised sub-labels
Upper Extremity	Wrist, Elbow, Arm
Lower Extremity	Leg, Knee, Hip, Ankle
Lateral Torso	Axilla, Breast
Posterior Torso	Back, Shoulder
Anterior Torso	Chest, Abdomen, Trunk
Palms/Soles	Hand, Thumb, Foot
Oral/Genital	Groin, Genitalia, Sacrum, Buttocks, Sacrum
Head/Neck	Neck, Chin, Face, Temple, Head, Forehead

Table 3.5: All the different labelling for the anatomical location of the lesion. Each label in the NHS data has been assigned to a category similar to the ISIC dataset.

3.7 Data Transformation and Augmentation

As mentioned in the data biases section the skin lesion images are taken under various conditions including angles, lighting, and distance from the skin lesion. While the variety of conditions will decrease the accuracy of results and hinder the detection of dermoscopic features, it is a requirement of the project.

One of the main challenges in melanoma detection is the visual similarity between normal and infected regions. Others are the presence of artefacts such as bubbles, hair and clinical marks[4]. These factors lead to low accuracy rates in traditional approaches. However, segmentation techniques can help overcome these challenges by removing these areas and isolating the melanoma from the rest of the image.

Skin lesion augmentation is especially vital because of the use of macroscopic images instead of dermoscopic images. This means there are various artefacts including hair, specular reflections, rulers, varying sizes, and shapes of the skin lesion. All of these can obscure the skin lesion and affect the accuracy of segmentation[115] and in effect feature detection.

By augmenting the skin lesion images using specular reflection removal and hair removal, the accuracy of feature classification methods can be improved[47].

3.7.1 Hair Removal

Hair artefacts in images can interfere with the recognition of handcrafted features and affect the performance of deep learning algorithms in melanoma detection[47]. Applying morphological operations such as image sharpening and segmentation techniques can remove hair artefacts from dermoscopic images[47].

Dull-Razor is an algorithm developed by Lee et al[54] and is frequently implemented with

Sharp-Razor[47] is a technique for detecting hair and ruler marks to remove them from images. This uses a multiple-filter approach including grayscale plane modification, hair enhancement, segmentation using tri-directional gradients, and multiple filters for hair of varying widths. This technique is shown to outperform existing methods.

3.7.2 Specular Removal

Specular reflection removal techniques are effective in improving the accuracy of melanoma detection[91]. A technique was proposed utilizing a partial differential equation to iteratively erode the specular component, removing the specular reflection[91].

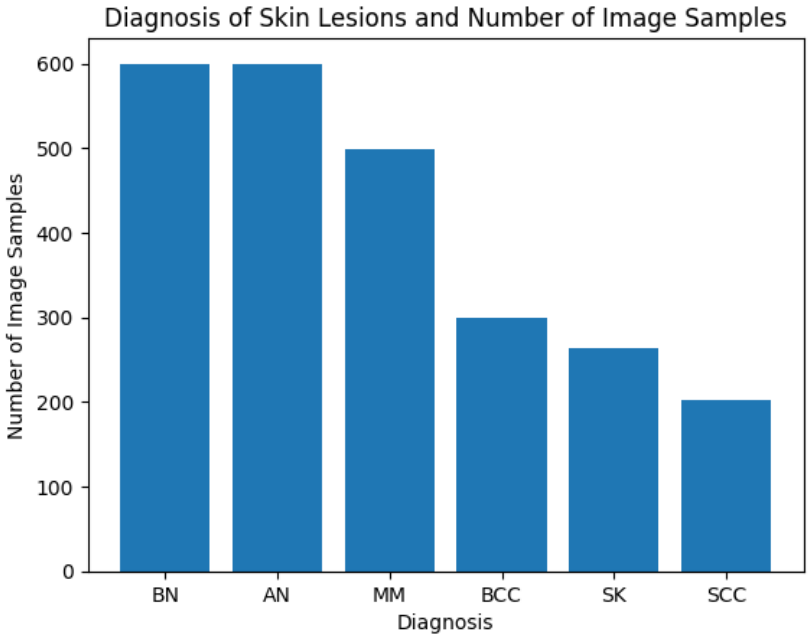


Figure 3.22: Number of image samples relating to the diagnosis of the image.

3.8 Conclusion

3.9 Dataset Statistics

The dataset has been analysed and modified accordingly, originally starting with a total of 2,500 images and it has been amended to 2,271.

As shown in figure3.9 the image data and diagnosis of the skin lesion there are several differences in this dataset compared with the ones described so far. There has been more of an attempt to balance the data so there are more equal samples of each. Furthermore, benign naevi have been split into benign naevi (BN) and atypical naevi (AN). There are images of seborrheic keratosis, which is more than any other public dataset currently available.

The age variation of patients described in figure3.9 demonstrates there are many younger patients included in the NHS dataset. This is substantially different from other datasets including ISIC that have mostly older patients. This demonstrates that there is an influx of younger patients regardless of them not being within the age group where melanoma usually develops. In both ISIC and NHS datasets the median the median is 60 years.

As shown in figure3.9 describing the location of the skin lesions and several image samples. There are more samples on the posterior torso (back) compared with other skin

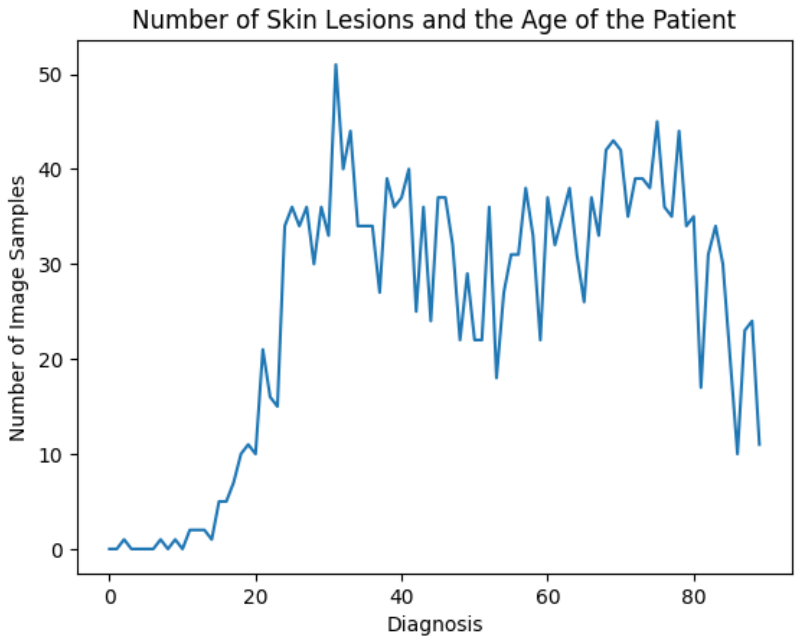


Figure 3.23: Age of patients and number of image samples.

lesions with any of the others. There are only a couple of samples for lateral, palms/soles, and oral/genital. This data was originally modified because it had a total of 28 descriptors, so they were grouped into 8 similar to the ISIC dataset. This can be seen in more detail in figure3.6.2.

Figure3.9 describes the number of image samples relating to the diagnosis and sex of the patients. Interestingly there are almost double the number of female patients being diagnosed for AN and BN compared with MM where there are slightly more male patients and BCC where there is almost double male.

Image samples described in figure3.9 demonstrate the age of patients compared with their diagnosis. Understandably, AN and BN which are neavus are from younger patients, while SK, SCC, and BCC appear in older adults. MM is primarily from patients at the age of 50 to 70.

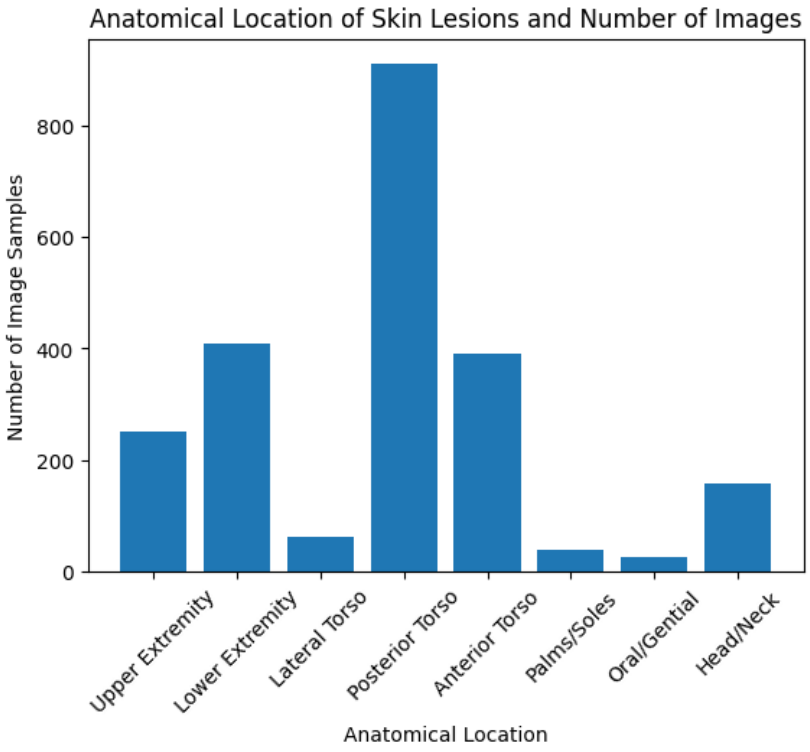


Figure 3.24: Number of image samples related to the location of the skin lesion.

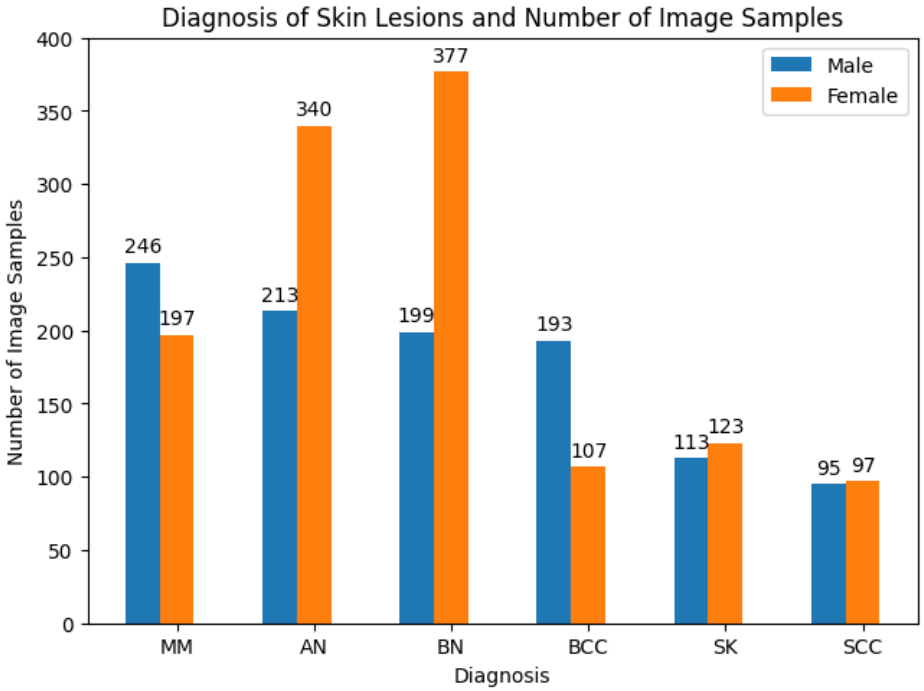


Figure 3.25: Number of image samples relating to the diagnosis and sex of the patients. There are more female than male patients.

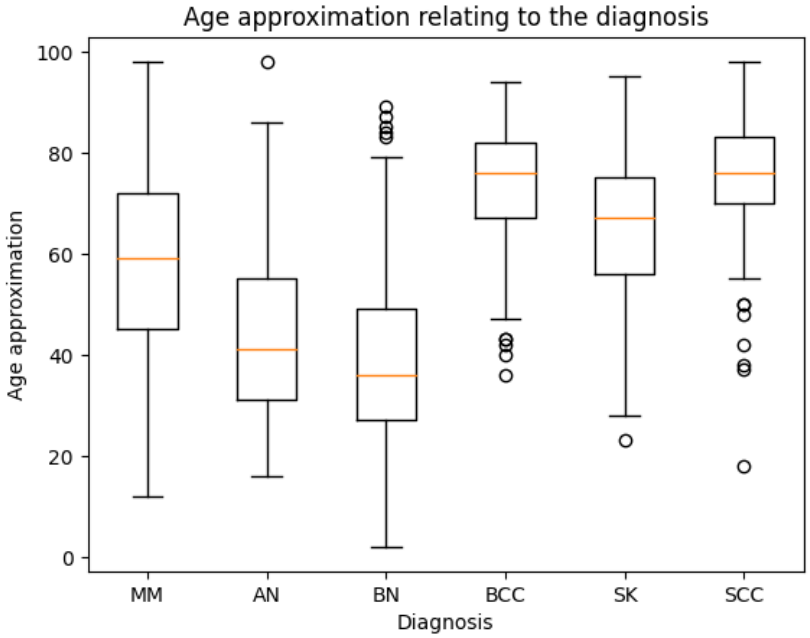


Figure 3.26: Boxplot describing the age of patients and the diagnosis.

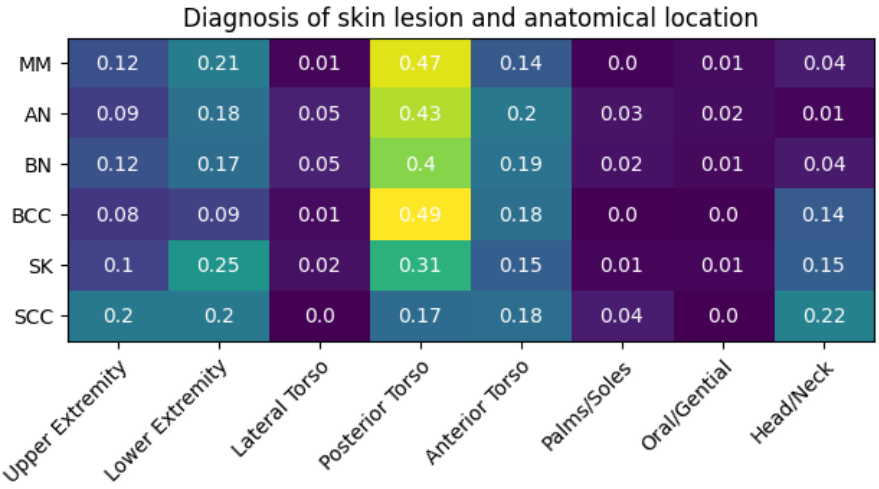


Figure 3.27: Number of image samples relating to the diagnosis of the image.

Chapter 4

Analysis of Explainability for the Detection of Melanoma

4.1 Introduction

This chapter contains an analysis of popular explainable AI (XAI) techniques called DeepSHAP and Gradcam. These techniques were compared to discuss whether their results are interpretable.

4.2 Background

Explainable AI (XAI) has gained significant attention in recent years because of increasingly more complex machine learning models in high-stakes decision-making processes in domains including healthcare, education, and public policy[8, 40, 29, 73, 106]. This issue was highlighted by the General Data Protection Regulation (GDPR and ISO/IEC 27001) has mentioned the concern for machine learning algorithms, mentioning the difficulty of implementation in the medical domain without adequate explanations[**empty citation**]. The public also has a preference for explainable systems**??**. Transparency, accountability and privacy are the most critical AI ethical principles[50] and they must be considered for use within sensitive domains including healthcare.

The lack of explainability in AI systems makes it difficult to evaluate the trustworthiness of algorithmic decisions, especially for the public and experts with little understanding of AI[29]. Issues could often arise with algorithms relating to data biases, such as the significant lack of data representing darker skin tones[81]. Without clinicians having an understanding of these issues, they might be misled into using incorrect diagnoses. This also highlights difficulty regarding accountability and whether the AI system or clinician would be blamed for any mishaps. Alongside this, there is a concern for parallel diagnoses[**empty citation**]. This referrs to AI systems that produce only a diagnosis with little to no explanation.

Without an explanation, the clinican cannot learn and attempt to understand why the results were met and in turn cannot utilise them. Considering the nature of clinical environments and that people's lives are at risk, algorithms need to produce explanations so that clinicians can intperpret and learn from results, but not depend on them.

Since highlighting the concerns of AI systems progress has been made in developing neural network architectures that are more interpretable. Techniques have since been developed to function with existing machine learning algorithms[37, 88, 79]. This is beneficial because many DNN architectures are the highest accuracy currently available[**empty citation**]. Other techniques[**empty citation**] involve extracting clinically relevant features such as ABCD rules or dermoscopic structures, followed by combining results into a diagnosis. Other issues with these technqiues are the current scepticism on whether these techniques are trustworthy[101, 85], and the concern they produce realistic but incorrect results[38]. Techiques such as LIME have been for use witthin

Some studies have described the use of explainable AI (XAI) models in healthcare[**empty citation**]. One of which shows that the confidence of clinicians is improved.

Lack of interpretability of AI systems has been identified as a challenge and these approaches are commonly referred to as "black box" approaches. This is because their inner workings are not visible and the system is

Some other interpretable techniques do not utilise neural networks. For example, Javier López-Labraca et al.[58] described an interpretable technique using multiple SVM models with colour and three dermoscopic structures (i.e., pigment networks, globules, and streaks). Bayesian fusion combines each model to calculate a diagnosis. Bayesian probability is a type of probability theory that uses probability distribution to estimate the values of unobserved variables. Bayesian fusion has comparable accuracy to neural network techniques[98]. Overall, results should be partially interpretable for use within clinical environments.

4.2.1 Dataset

Comparisons were made using the ISIC 2019 dataset because it is the largest and most robust public dataset currently available regarding melanoma detection.

4.3 DeepSHAP

DeepSHAP (Shapley Additive exPlanations) is a method designed to offer insights into the decision-making process of machine learning models, specifically deep neural networks (DNN). DeepSHAP is an extension of the DeepLIFT algorithm and is based on the concept of Shapley values that are derived from cooperative game theory. The method aims to estimate the importance of input features for a given decision by comparing the activations in the network for a given input against the activations caused by a reference input. In Turn, DeepSHAP is particularly effective

The method is particularly effective for explaining the performance of deep learning models in medical decision support systems[**empty citation**]. It has been shown to highlight information relevant to the decision-making process. This is more effective than layer-wise relevance propagation (LRP), local interpretable model-agnostic explanations (LIME), and DeepLIFT.

In the field of healthcare, DeepSHAP is applied to predict and explain non-communicable diseases (NCDs). In explanations for individual predictions and a case study detecting the progression of Alzheimer's.

Although explainable algorithms have seen some use within healthcare, there is no evidence of their current use within dermatology.

DeepSHAP (Shapley Additive exPlanations) is a game theoretic approach designed to explain models during training by visualising features related to the classification. It explains the individual predictions in machine learning models using Shapley values that measure the contribution of each feature to the contribution of an outcome[1].

4.3.1 Summary

4.4 Grad-Cam

4.4.1 Summary

4.4.2 Tree ensemble methods

4.5 Bayesian Network Approach

4.6 Conclusion

Chapter 5

Segmentation and Border-line cut-off

5.1 Introduction

5.2 Background

Segmentation plays a crucial role in correctly identifying melanoma. The accurate recognition of melanoma is a challenging task due to the contrast between lesions and skin, including visual similarities between melanoma and non-melanoma skin lesions[56]. For example, classification tasks are imporved by separating the lesion from healthy skin ensuring that only relevant sections of the lesion are classified[4, 19]. The goal of producing segmentation techniques is to simplify and transform the representation of images into a more meaningful form for analysis, which is essential for precise representations of melanoma lesions[59]. This essentially leads to a more effective diagnosis by identifying key features that indicate malignancy[5]. Overall segmentation is vital for the accurate and reliable analysis of melanoma.

Accurate segmentation also plays a for the analysis of ABCD rules[53]. Especially for the analysis of border features[71, 48] including convexes and indents. The irregularity of borders is a key feature in distinguishing melanoma from benign lesions, and the exact identification of irregular borders from melanoma skin lesions is clinically significant[70]. An accurate border cut-off is also important for the analysis of asymmetry and colour which both classify the colour of the skin lesion. An accurate border is also essential to analysing lesions by preventing skin from accidentally being calculated in the process, ruining the accuracy. It's important to note that many of the ABCD rules algorithms are statistical and highly sensitive to incorrect segmentations.

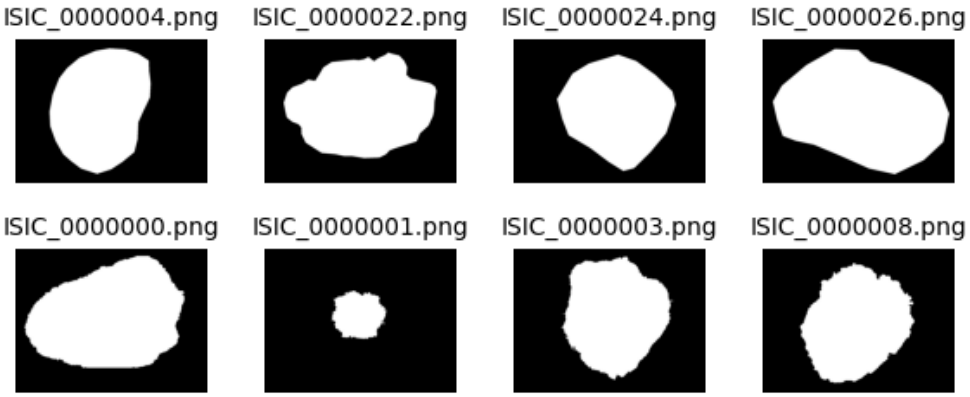


Figure 5.1: This figure shows examples of different styles of segmentation in the ISIC 2018 dataset with approximate segmentation masks on the top row and expert segmentation masks on the bottom.

5.2.1 Significance of Border cut-off

The border cut-off sometimes referred to as an expert border is essentially a precise border that captures the edges between the skin and skin lesion. This is shown further in figure 5.7 from the ISIC 2018 dataset that has a mixture of approximate and expert border types.

Many of the deep learning algorithms produce approximate borders; rough area around the lesion with no border cut-off for the lesion. The ABCD rules technqiues rely on border cut-offs. For example, colour analysis relies on removing any remaining skin colour and border relies on indents and convexes that are not present in approximate borders.

Essentially, specifcially for the analysis of ABCD rules the use of deep learning (SegNet, Unet) is not effective enough for their analysis. Therefore, statistical technqiues (Otsu, LBPC) are tested to adjust the border as a joint technique.

5.2.2 Methodology

In evaluating the effectiveness of trained models, several training parameters are commonly used. These parameters serve as quantitative measures to assess the performance of models. The following parameters are utilised:

- Loss: The loss function is a parameter used to measure the inconsistency between the predicted segmentation and the ground truth. Loss is typically minimised as the model is trained through the optimisation algorithm including Adam or stochastic gradient descent (SGD).
- Recall: The recall function otherwise known as sensitivity evaluates the ability of the model to correctly identify instances within an image. For example, it is calculated as

a true positive to the sum of true positives and false negatives. A high recall value means that the model is effective at capturing the relevant objects in the image.

- Accuracy: The accuracy is a fundamental metric that measures the overall correctness of the model's classification. It is calculated by the number of correct predictions to the total number of number of predictions. Accuracy tends to be ineffective in analysing unbalanced datasets.
- The precision quantifies the accuracy of the model's positive predictions. It is calculated as the true positive to the sum of true positives and false positives. A high precision indicates that the model has a low risk of producing false positive predictions.
- Dice Coefficient: The dice coefficient that is also known as F1 score is a metric that combines precision and recall into a single value. It is calculated as the harmonic mean of precision and recall. This provides a balanced assessment of general model performance in segmenting images.
- Jaccard Index: The Jaccard index or Intersection over Union (IoU) measures the similarity between the predicted segmentation and ground truth. It is calculated with the intersection of the predicted and ground truth regions to the union, providing the model's segmentation accuracy.

These parameters are a collective evaluation of the SegNet model's performance in image segmentation. These allow researchers to assess the effectiveness of the models and how they might function in real-world scenarios.

5.3 Deep Learning Segmentation Algorithms

5.3.1 Semantic Pixel-Wise Segmentation (SegNet)

Semantic Pixel-wise segmentation termed SegNet is a deep convolutional encoder-decoder architecture designed for the automatic segmentation of images. It was originally developed in 2015 by Badrinarayanan et al.[14] and has shown promising results for various segmentation tasks including those in the medical field. SegNet has been used in a variety of applications in the medical field including dental imaging[52], liver tumour segmentation[74], and many others. It is known for its efficiency in terms of memory and computational time and is known for being more memory efficient than other architectures including U-Net[63]. Further developments have been made in upgrading SegNet into different versions including Bayesian SegNet[30], and transfer learning using VGG-SegNet[**daniel2022**], and ResNet-SegNet[**mohd2020**], with differences in the encoder and decoder layers.

The engine consists of an encoder network with identical 13 convolutional layers. The idea of SegNet is to perform pixel-wise classification by assigning each pixel in an image to

a specific class or category. This is achieved through a deep convolutional encoder-decoder architecture, which allows for robust semantic pixel-wise labelling.

Semantic pixel-wise segmentation (SegNet) is a machine learning architecture utilizing a deep, fully convolutional neural network (DCNN). This network requires training from ground truth and pre-segmented images for automatic segmentation. SegNet consists of encoding layers, decoding layers, and a pixel-wise classification layer. The encoder layers consist of 3×3 convolutions (including batch normalization and ReLU), and pre-trained filters for classifying features. After some convolutions, the data is down-sampled using a 2×2 pooling layer. Next, decoding layers consist of up-sampling, followed by 3×3 convolutions. Finally, the pixel-wise classification uses a softmax layer to represent each pixel between 0 and 1 based on the previous layers, generating a segmentation mask.

The optimisation model Adam is used and was designed to replace stochastic gradient descent (SGD) because it is generally better than other models and has a faster computational time.

After training the model using parameters and 100 epochs the trained model has the following metrics, shown in figure 5.2. In the diagram blue represents training data making up 80% of the overall data and orange is the validation data making 20%. All the training parameters increase to above a certain percentage, and although the validation appears to have stopped increasing it isn't decreasing. If it was decreasing it would demonstrate overfitting.

Here are some examples of the SegNet model described in figure 5.3. SegNet appears to capture the area of the skin lesion with high accuracy, but when compared with the expert borders in the images it fails to capture these features.

Results in figure 5.3 are generated from the architecture using the ISIC 2018 dataset split into 80% training and 20% validation images. The accuracy of locating the lesions is 85%. However, it represents the border cut-off between skin and skin lesion accurate to the dataset but inadequate for using the ABCD rules. Finding the border cut-off is vital for measuring ABCD rules[71].

5.3.2 Unet

U-Net or U-shaped neural network is a full convolutional neural network (FCN) architecture built for image segmentation tasks. It is an encoder-decoder architecture designed for semantic segmentation tasks, it is especially useful for medical image analysis. One advantage of this model is the ability to use high-resolution images and produce accurate segmentation maps.

The model consists of two 3×3 convolutions (unpadded convolutions), each followed by a rectified linear unit (ReLU) and a 2×2 max pooling operation with stride 2 for downsampling. After each downsampling the number of features is doubled. Then when upsampling the features are halved alongside 3×3 convolutions, each followed by a ReLU. The final layer consists of a 1×1 convolution that is used to map each 64-component feature

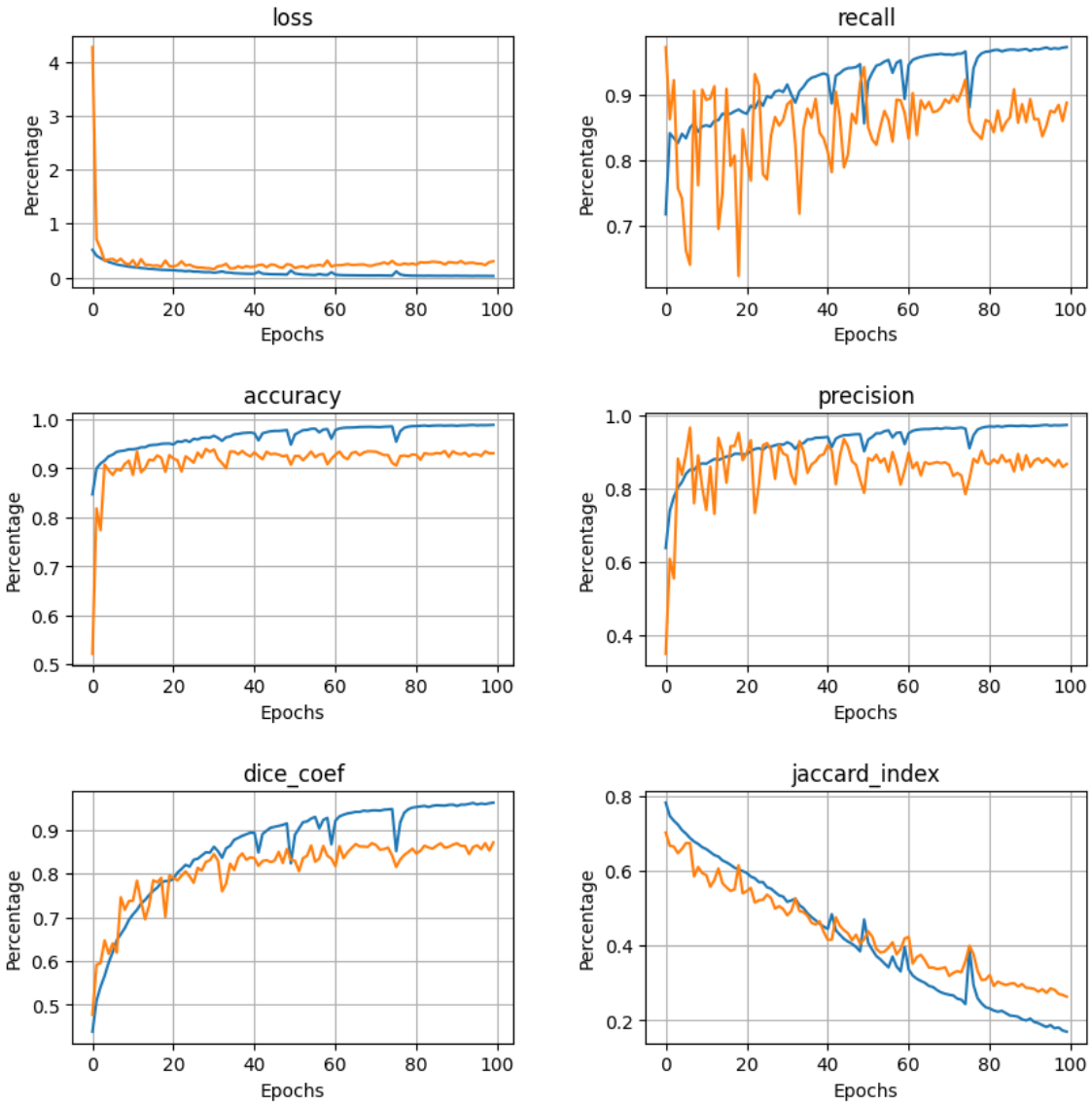

loss
Percentage
Epochs
recall
Percentage
Epochs
accuracy
Percentage
Epochs
precision
Percentage
Epochs
dice_coef
Percentage
Epochs
jaccard_index
Percentage
Epochs

Figure 5.2

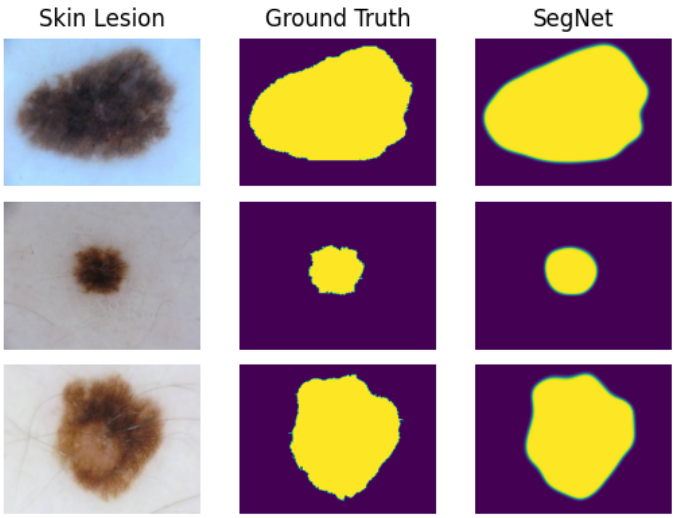


Figure 5.3: SegNet model segmentation compared with the skin lesion, ground truth, and segmentation.

vector to the desired number of classes.

5.4 Border-line Cut-off Segmentation Techniques

5.4.1 Otsu Threshold

Otsu threshold is a versatile automatic image thresholding technique meant to separate each pixel between two classes of foreground or background. One of the benefits of this method is that it does not require any training data. The equation 5.4.1 (within-class variance) describes splitting weights of $w_0(t), w_1(t)$, which are the probabilities divided by the threshold t, between 0 and 255. Furthermore, σ_1^2 and σ_0^2 are variances of these two classes. The class probability w is computed from the histogram in figure5.4, which is an intensity histogram describing the colour distribution in an image. Measuring the values above and below the generated thresholds splits the image into two classes.

$$\sigma_w^2(t) = w_0(t)\sigma_1^2(t) + w_1(t)\sigma_2^2(t) \tag{5.1}$$

The histogram was split into two segments with the threshold t of 138 and the corresponding pixel locations to the histogram segment the skin lesion into two classes. Image morphology closing was applied to fill gaps that the threshold missed. On other occasions, the segmentation missed the skin lesion because of a similar colour between the skin and the skin lesion. It might be beneficial to combine Otsu with SegNet to improve its accuracy

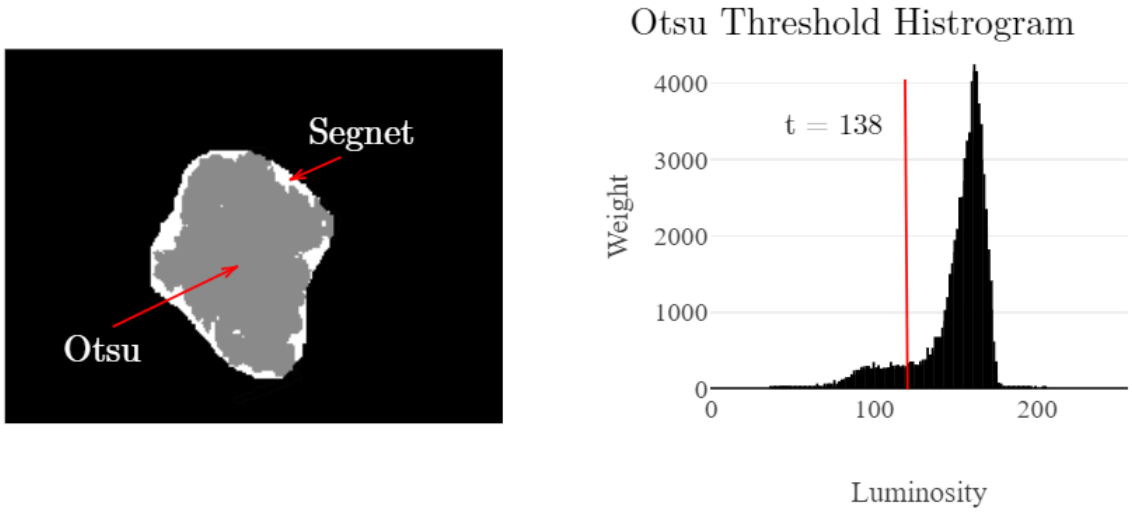


Figure 5.4: Otsu thresholding alongside ground-truth mask, where grey Otsu and white is SegNet. The bar chart shows the histogram with an otsu threshold of 138.

while producing a border cut-off. Figure5.4 describes the difference between otsu and SegNet.

5.4.2 Local Binary Pattern clustering (LBPC) Segmentation

Local Binary Patterns (LBP) is a texture descriptor commonly used for augmenting the image improving classification accuracy[71, 48]. First, equation5.2 calculates each pixel, where p (equal to 8) is the number of neighbouring pixels compared to the centre of c, and the radius of r from the centre. Next, shown in equation5.3 each value is subtracted counter-clockwise with the centre value and compared to function S where each $gp - gc$, if more than or equal to 0, is equal to 1, and less than 0 is equal to 0. Next, add corresponding values equal to 1 of gp together, changing the centre value, ignoring values of 0. Next, applying a Gaussian kernel of 13-pixel iterations and a standard deviation of 3 removes smaller features that interfere with the segmentation. Finally, applying k-means with a value of 2 subtracts the greyscale and segments the skin lesion from the skin.

$$LBP(gp_x, gp_y) = \sum_{p=0}^{P-1} s\big(gp - gc\big)2^p \tag{5.2}$$

$$s\big(x\big) = \begin{cases} 1, & x \geq 0; \\ 0, & \text{otherwise.} \end{cases} \tag{5.3}$$

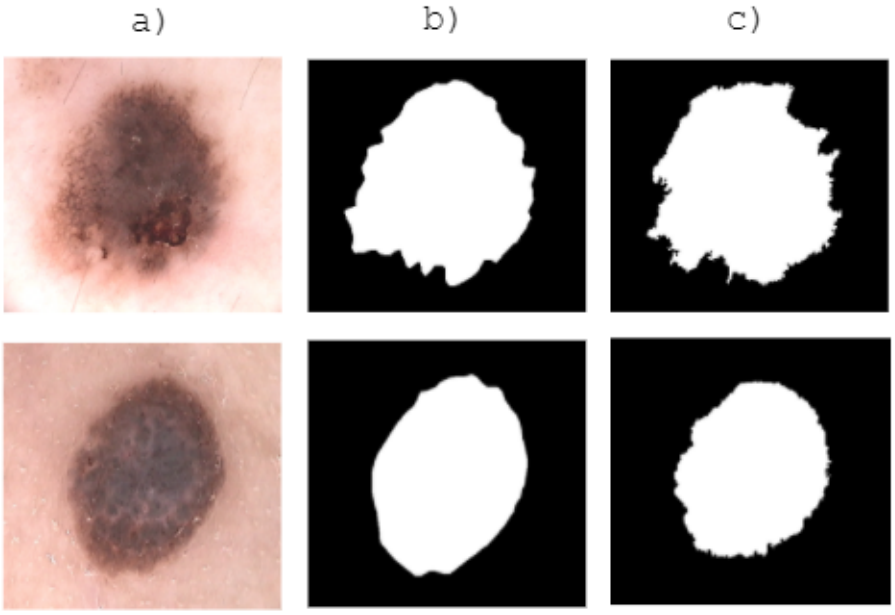


Figure 5.5: Local Binary Pattern Clustering (LBPC) showing the a) original image, b) ground-truth, and c) LBPC. LBPC successfully exaggerates the border cut-off on the skin lesions with regular and irregular borders

Figure 5.4.2 demonstrates the segmentation of two skin lesions, one with an irregular border and another with a regular border. LBPC is applied to both skin lesions, followed by Gaussian blurring and morphology closing to remove dots. The result is an improved border cut-off compared to the ground truth in the Ph^2 dataset with more corners and ledges. This technique will improve accuracy for measuring border irregularity[71].

Validating LBPC is not expected because the goal is to exaggerate the border to improve the classification process of ABCD rules, which it does successfully[71, 48]. For example, the segmentation might not match dataset segmentations but is still essential to classifying ABCD rules. Furthermore, many datasets lack expert border segmentation, an accurate border cut-off between the skin and skin lesions, so comparisons are not always possible.

5.4.3 Issues

One issue with this technique is the overreliance on the size of the skin lesion. For it to function properly it needs to have a certain amount of skin and skin lesion for the segmentation to be possible. This leads to some concern when it comes to melanoma which is frequently larger and takes up more space in the image. This means that the accuracy of LBPC is in question for larger skin lesions. Here are some examples where the algorithm

Figure 5.6

failed:

5.4.4 Results

Overall the accuracy of the techniques demonstrates that SegNet is the most reliable technique. However, comparing the techniques in **??** we can demonstrate that it produces a smudge effect and fails to capture the border cut-off from the skin lesion, but it is successful at finding the location of the skin lesion.

Both statistical models of LBPC and Otsu threshold generated an accurate border cut-off between the skin and the skin lesion As previously mentioned, measuring the border cut-off and exaggerating irregular borders are helpful when calculating the ABCD rules.

It might be beneficial to combine SegNet and LBPC using SegNet to find the skin lesions' location, followed by adjusting the border cut-off using LBPC. A similar technique using the Otsu threshold and Segnet is described by Riaz et al.[78].

5.4.5 Results

5.5 Experimental Results

This section includes a simple border analysis technique called fractal box counting to assess the benefits of using different segmentation algorithms with accurate border cut-offs.

To prove the usefulness of segmentation techniques with an accurate border cut-off a technique developed by Ali[6] is implemented that utilises machine learning with extracted data including Zernike moments, fractal box-counting, and convexity measurements. Fractal box-counting is used to measure the irregularity of the border.

The fractal box-counting technique is a commonly employed technique for analysing fractal properties. It involves dividing a fractal object or pattern into a grid of equally sized boxes and counting the number of boxes that contain a portion of the fractals. The process is repeated with different box sizes until the boxes until the relationship between the box sizes and the number of boxes is analysed determining the fractal dimension[39]. Essentially a more complicated border with corners and convexes will have more boxes and therefore a higher fractal score, than for example a border with smooth corners and edges which has a lower score. This should provide some evidence of the usefulness of an accurate border.

5.5.1 Issues

Although these techniques are massively accurate they are trained and tested against datasets (ISIC 2018) containing a mixture of approximate segmentation masks and expert

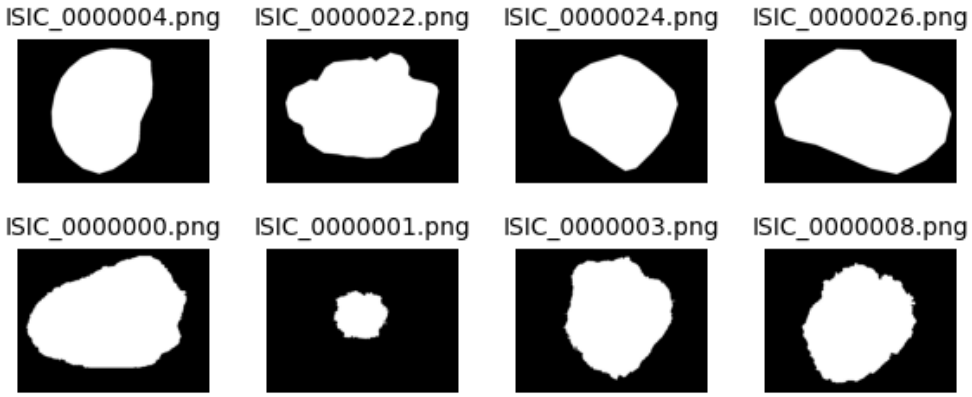


Figure 5.7: This figure shows examples of different styles of segmentation in the ISIC 2018 dataset with approximate segmentation masks on the top row and expert segmentation masks on the bottom.

segmentation masks.

As mentioned in the previous section the ISIC 2018 dataset contains a mixture of approximate borders and expert borders. The statistical algorithms (otsu, LBPC) generate an accurate border cut-off and would have better accuracy for expert borders, but worse for approximate borders. The opposite is true for deep learning algorithms (U-net, SegNet) where they produce approximate borders and should ideally be tested against approximate borders.

Considering these problems to better assess the quality of algorithms the ISIC 2018 dataset could be split into different border types. This still has its issues because the border cut-off is subjective and where the skin lesion and skin end is still up to debate. The reason for extracting the border cut-off from the skin lesions is for better analysis of ABCD rules. So, both border types are assessed using fractal box counting, where the value should increase with more complex borders.

Figure 5.7 shows approximate borders are a rough estimation of skin lesion area and expert segmentation masks fit the skin lesion tighter representing border features. The mixture of images results in produced borders from many deep learning techniques having an inaccurate border cut-off border, whilst appearing accurate when tested against datasets. The ISIC 2018 dataset has a mixture of approximate segmentation masks and expert borders, which is not enough to train deep learning algorithms. Furthermore, the images are also shrunk as part of the deep learning process, which in turn loses smaller features that are significant when analysing borders. Various hybrid techniques have been developed using statistical algorithms active contouring-based segmenatation[78], LBPC and others for border adjustment including u-otsu and edge-imfill. Specifically, u-otsu has been previously used to adjust the borders of melanoma to create expert borders[**empty citation**].

As shown in figure 5.7 expert borders are tighter to the area of the skin lesion and approximate borders are the area of the lesion. Images have either approximate borders or expert borders and no others in the ISIC 2018 dataset, measurements are likely skewered because of the variation in types of borders.

The segmentation algorithms encountered some issues, whilst the best of the techniques was SegNet with an 85% accuracy when relating to the ISIC 2019 dataset. However, as previously mentioned the segmentation masks have poor border cut-off stunting features that are useful for finding border irregularities.

In contrast LBPC and U-Otsu algorithms effectively identify the border cut-off of the skin lesion, which isn't properly represented in the ISIC 2019 dataset. But, it sometimes fails to find the skin lesion or does not detect anything.

Both techniques appear to have downfalls making them less effective for use for analysing ABCD rules. It would be beneficial to find the approximate area of the skin lesion using SegNet and followed by LBPC to find the border cut-off.

5.6 Joint Neural network and statistical model approach

Combining both SegNet and LBPC improves the accuracy.

5.7 Results

Chapter 6

ABCD rules, and Dermoscopic Structures

6.1 Introduction

This chapter is a discussion of the most popular ABCD (Asymmetry, Border, Colour, and Dermoscopic Features) algorithms including their implementation, and updating the algorithms. They are compared using the PH2 dataset and updated relating to their accuracy. Surprisingly, many of the ABCD rules techniques were originally tested for whether they effectively find melanoma and not individual features. So, this will be the first time some of these techniques have been tested and documented.

6.2 ABCD Rules Data Extraction Techniques

The automatic detection of ABCD rules for the detection of ABCD rules has warranted extensive research and development[7]. The ABCD rules stand for Asymmetry, Border irregularity, Colour variegation, and Diameter greater than 6mm, and have been fundamental framework for the clinical diagnosis of melanoma. Sometimes Diameter is changed for Dermoscopic structures because measruing the size of the lesion is not always possible when capturing an image. Furthermore, demroscopic structures provides further insight into diagnosing melanoma[**empty citation**]. Another useful change made to the ABCD rules is E for evolution, showing that it changes overtime, datasets however do not support this. Although the diagnostic procedures ABCD rules is frequently used in the medical environment it has limitations in detection of small melanomas and those with resular shape and homoheneous colour[12]. As a result, automatic classification algorithms for ABCD rules has become more prevelant[46].

The benefits of implementing a CAD system for the automatic detection of ABCD rules is the classification of exisiting diagnostic procedure in which clinicians are already aware

Figure 6.1

of how it fundementally works. Another benefit is the automatic labelling of data, where it would take considerable time to input this diagnostic data into a server for further analysis, where the automatic system could generate results and the clinician only needs to check whether these results are correct. Other benefits include improving objectivity of results, where normally ABCD rules is considered subjective, meaning it can be utilised in a variety of ways. This ultimietly makes comparisons very difficult between lesions, but automatic detection can improve the objectivity, making hospital wide comparisons easier. Check out the previous chapter discussing the PH2 dataset and the data subjectivity on page48.

6.2.1 Asymmetry Techniques

Asymmetry analysis is a fundamental component in the early detection of melanoma because it often exhibits asymmetric shapes[5]. Meaning that the shape, colour and, texture match asymmetrically more often in benign lesions. For example, as melanoma grows the central area begins to waste away leaving a hollow area covered by thin skin, showing dermoscopic features. As it grows the edges become more irregular producing an uneven shape often relating to irregular borders and asymmetrical shapes. Diagnostic procedures have been developed to detect these unique characteristics.

Figure6.2.1 demonstrates the difference between the updated asymmetry algorithm using super pixels compared with the original

the asymmetry algorithm using super pixels decreases the median and range, compared with the original method where the pixels are closer together

6.3 A Novel Asymmetry detection technique using Bi-Fold, 3D Euclidean distance, and Superpixels

This section describes a novel machine-learning technique for the automatic detection of melanoma

6.3.1 Image Transformation

Transforming the skin lesion image is important to prevent scale variance in the algorithms. For example some of the images have a smaller skin lesion in upper corners and others take up the entire image size. For adequate comparisons the images must be transformed to match sizes. THis ensures the images are being compared fairly.

To transform the image it is converted to gray scale and a threshold is applied finding all non-zero pixels. This is followed by a bounding rectangle around the area being assigned. After this the assigned area from the original image is cropped.

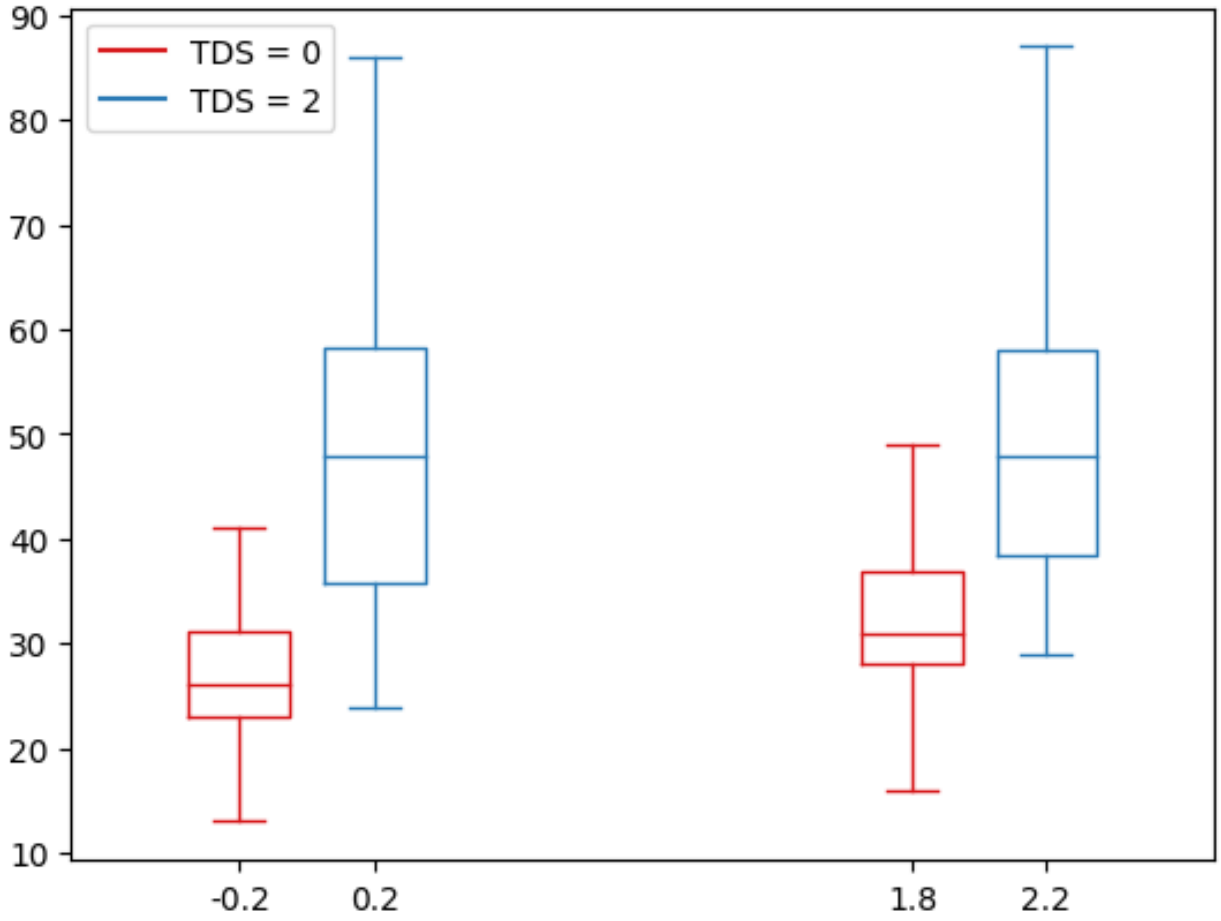


Figure 6.2

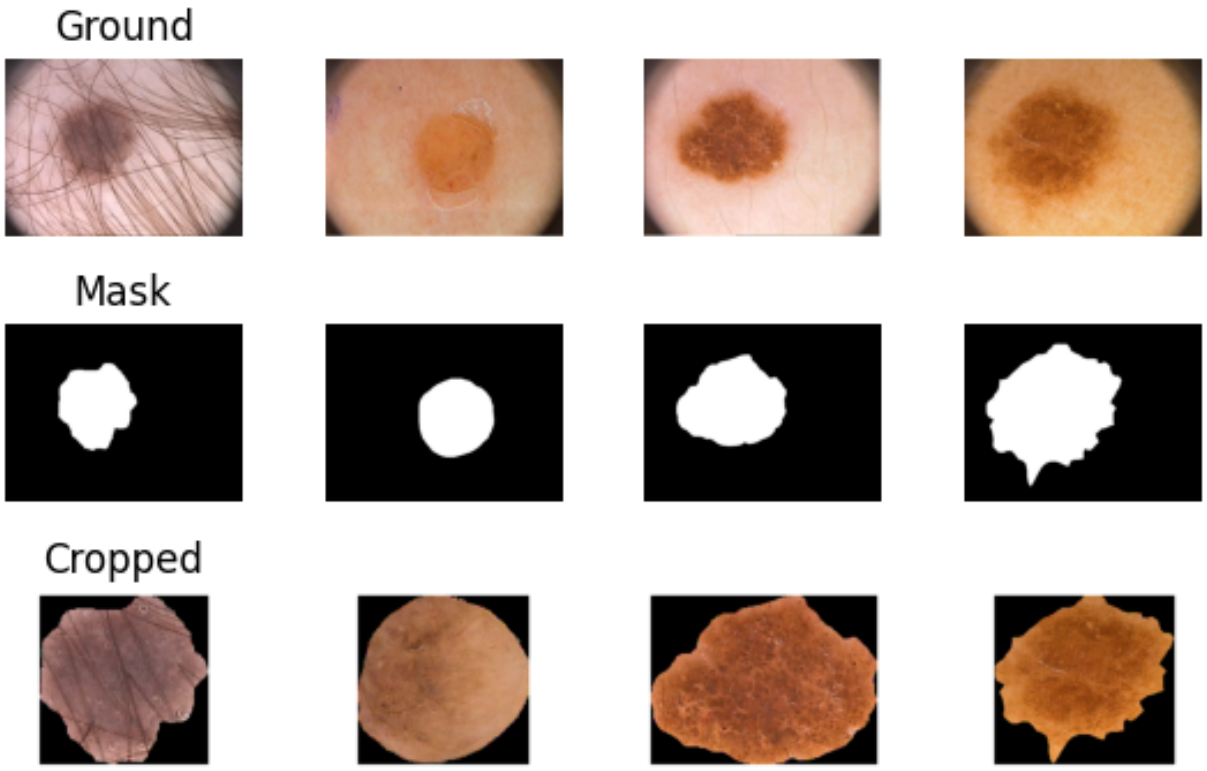


Figure 6.3: This figure shows some images from the PH2 dataset after being masked, cropped, and rotated using bi-fold.

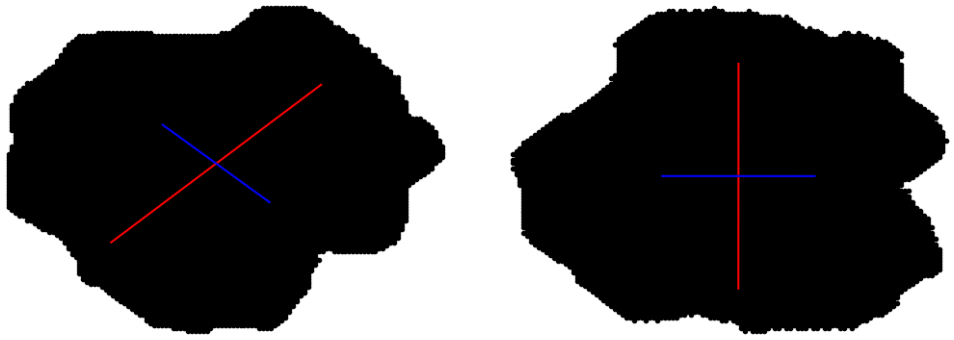


Figure 6.4: This diagram shows image IMD003 from the PH2 dataset after calculating Bi-fold, followed by rotating the image to match the rotation defined using moments of inertia.

The first line in figure 6.3.1 shows the skin lesion image, the second shows the segmentation mask, and the third shows the images after being cropped and rotated using bi-fold.

6.3.2 Bi-fold

Bi-fold is a diagnostic procedure designed to support the recognition of melanoma by drawing a line down the middle of the skin lesion and comparing the two halves to confirm whether the sides match (considering the difference in shape, colour, and texture). Using this horizontally and vertically calculates whether the skin lesion is possibly malignant with a score between 0 and 2. Calculating with Total Dermoscopy Score (TDS) alongside the other ABCD rules including asymmetry, border, colour, and diameter calculates the likelihood of malignancy. Dermatologists frequently use bi-fold due to its simplicity, but it can be subjective to the original observer and time-consuming when managing large numbers of skin lesions. Therefore, automating techniques is beneficial to clinicians and can improve the objectivity of results.

To initiate the classification of skin lesions a technique called bi-fold is applied involving folding the skin lesion in half vertically and horizontally and a comparison of their respective dimensions. While the original technique was designed only to assess the lesions' shape, it's been utilized to account for colour and texture as well. The centre and orientation are determined by calculating its moments, where the centre is (m10 / m00, m01 / m00) and phi is 0.5tan (2m11)/ (m20 -m02).

6.3.3 Shape analysis

Analysing the shape of the skin lesion

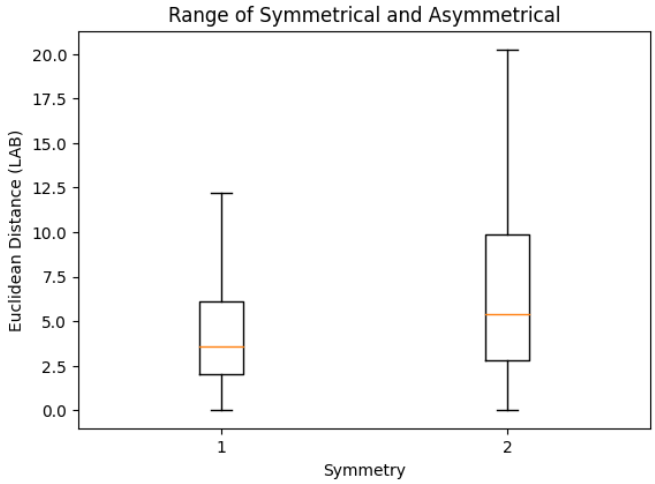


Figure 6.5: This diagram is a summary of the PH2 dataset after using bi-fold, a Euclidean distance of colour. The value on the right would be a threshold.

6.3.4 3D Euclidean Distance

Next, the lesion is partitioned into a 20 by 20 grid centred on the mentioned centre point, and the average of each region is computed. This is followed by finding the matching region on the perpendicular area from the centre of the skin lesion and comparing the colour distance between the two. Distance is measured using the LAB colour space and a 2D Euclidean distance of A and B, removing L (luminosity) to eliminate light variation. Once compared, all compared regions are obtained, and they are plotted onto a graph. If over half of the values are above a threshold of 6, then the lesion is asymmetrical.

The diagram shown below in figure6.3.4 is a compilation of all the images within the PH2 dataset showing the threshold range after applying bi-fold, euclidean distance of colour, but before applying the threshold. As can be seen, a threshold of 6 covers all of the symmetrical values, but still roughly covers half of the asymmetrical values. This demonstrates that the technique produces many false positives when regarding asymmetrical values. Essentially, the symmetrical skin lesion has a smaller area and the asymmetrical lesion has a larger area, but both remain in the same zone and therefore splitting the data only using a threshold holds poor results. Furthermore, there are a lot of fliers and the threshold does not adjust according to these values. See the graph below:

To improve the accuracy of the algorithm some changes need to be made based on the previous statements. First will be superpixels and next is k-means.

6.3.5 Superpixels using Simple Linear Iterative Clustering (SLIC)

Superpixel is an algorithm for grouping pixels into a grid format, but with flexible borders that can adjust to regions with similar features. Unlike the original technique averaging specific squares in a grid[46], they are segmented related to colour, texture, and other properties. The reason for using this technique is to increase boundary adherence and to group features that might otherwise be split into separate groups. This overall improves the accuracy of the algorithm.

This technique uses a simple linear iterative clustering (SLIC) algorithm and was first introduced by Achanta et al.[3]. The technique combines both k-means and graph-based segmentation. Firstly you define the desired number of superpixels as k and the approximate size of each superpixel as S, which is usually $S = /square(N/k)$ and N is the number of pixels in the image. Secondly, for the centre of each cluster, a search space is assigned to the cluster. For each group, you measure the spatial distance which is the Euclidean distance between each pixel and the cluster center. Each pixel is assigned to the cluster with the nearest centroid. The cluster centres are then recalculated by taking the mean colour and position of the pixels assigned to each cluster. Followed by new pixels being assigned to the centroid relating to Euclidean distance. This process is repeated depending on the number of iterations as i are assigned. From this point, each pixel is assigned to a cluster.

The image in ... demonstrates the usual average and the new averages based on superpixels and the changes in values. Areas that are lighter in colour appear to have a lower value and darker appear darker.

Using the thresholding method for classification we can already see the accuracy has been improved with a threshold of ...

6.4 Experimental Results

The goal of this experiment is to improve the accuracy of the asymmetry bi-fold technique described by Ihab S. Zaqout et al.[113]. Initially, the skin lesion is split into a 10 by 10 grid and converted into the LAB colourspace. Next, a line is drawn through the middle horizontally and vertically. Measuring the Euclidean distance from the centroid, locating the closest opposite patch of colour finds the parallel square. Subtracting the squares generates a score for each value, the closer to 0, the more similar the colour. These are then removed from the list to prevent them from being selected a second time. If half the results are over a specific threshold, it is considered asymmetrical in colour, otherwise considered symmetrical. The aim is to make a 10 by 10 grid, but instead of averaging squares, superpixels reduce data redundancy in the grid, allowing for a less complex algorithm and improving accuracy. The clustering method k-means partitions each pixel to its nearest most similar centroid relating to colour. Next, it generates a superpixel that represents the average colour of that area. The diagram6.4 demonstrates different borders when changing the C for compactness, where 100 generates a square grid similar to the original technique. The border becomes

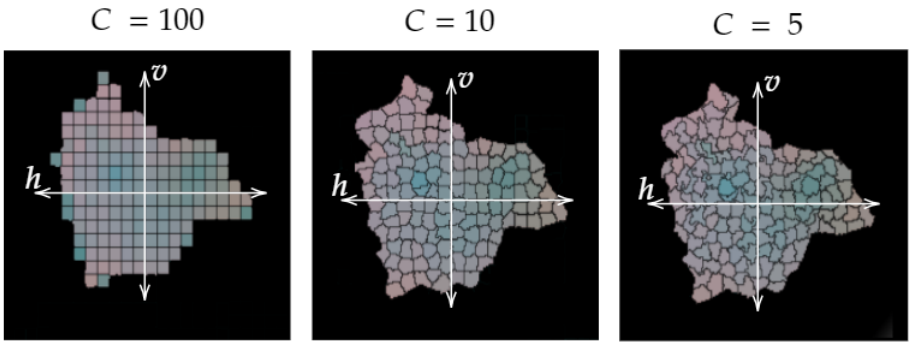


Figure 6.6: This diagram shows the skin lesion split relating to superpixels instead of averaging squares.

more flexible as the compactness value decreases.

Each parallel square on the vertical and horizontal axes measures similarity using a three-dimensional Euclidean distance in the LAB colour space. For example, the perceivable difference of colour to the human eye is a three-dimensional Euclidean distance of 6[67]. Using similar logic, a value of 20 is the threshold, where any value over that amount is considered asymmetrical in colour. Next, each square is compared with its closest parallel square (relating to the line through the centre defined by the bi-fold) and removed from an array after being compared. The next improvement is to generate a unique threshold for the significance of each square. For example, using superpixels with the compactness of 10 has an accuracy of 61% with the PH^2 dataset compared to the original 59.5%. This approach demonstrates that a flexible border that considers features is more effective than averaging squares.

There is a correlation in colour differences between the inner and outer edges because melanoma typically expands outwards, creating an abnormal border. This information specifies that the statistical model accuracy could be improved by increasing the threshold for the outer edges and decreasing it for the inner.

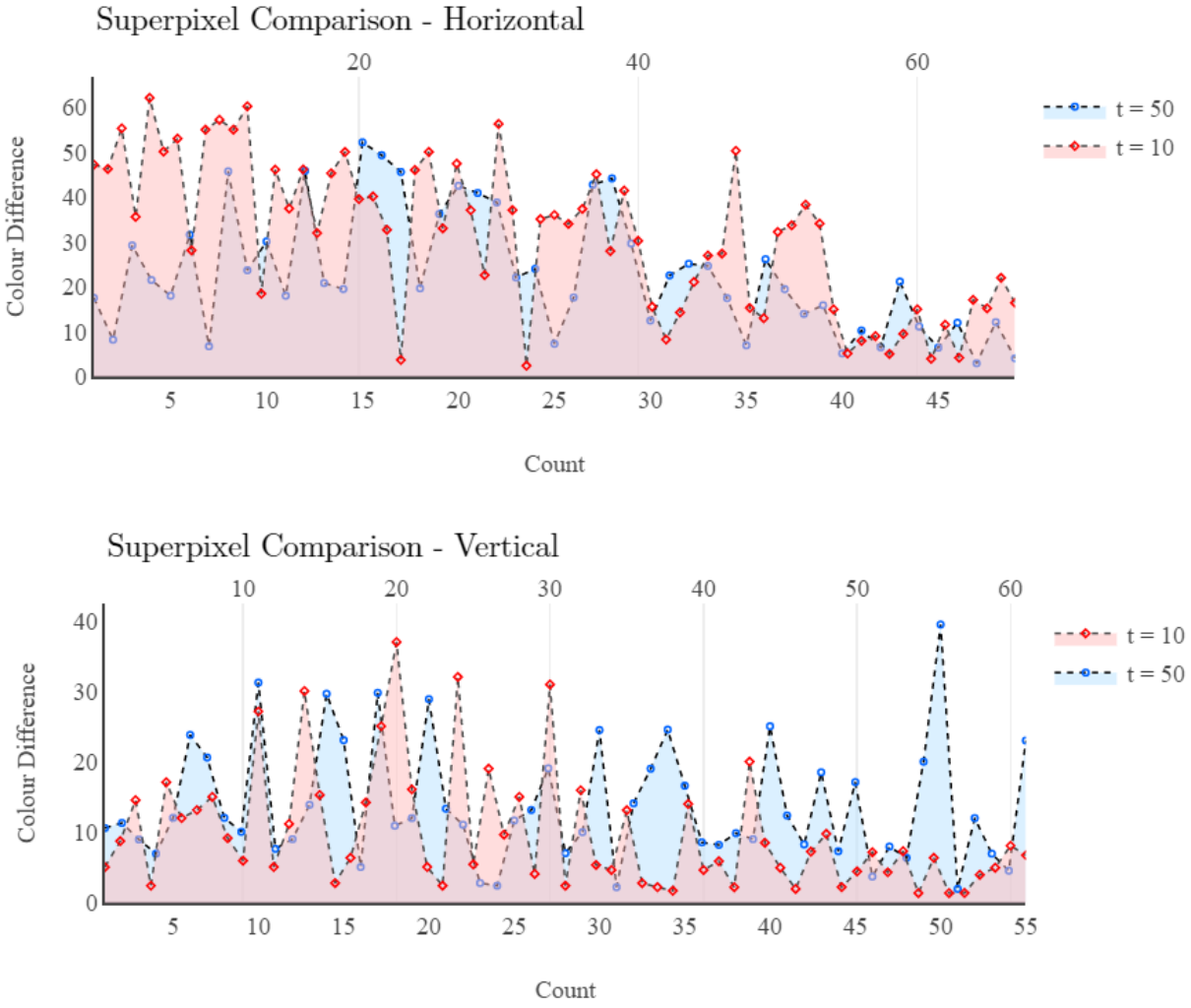


Figure 6.7: This diagram shows the difference between averaging squares and using superpixels, with the threshold of 10 implying curves and 50 being squares. The horizontal colour difference is improved, making it more likely to be seen asymmetrical. The vertical comparison is roughly the same, except for removing a false positive of 40.

6.5 Border Detection Using Zernike Moments, Fractal Box-Counting, and Convexity

6.6 A Novel Colour Analysis Approach using Colour Ranges, and SVM

6.7 Dermoscopic structures

6.8 Results

6.9 Conclusion

Chapter 7

Combined ABCD Rules and Dermoscopic Structures using Bayesian Network

7.1 Introduction

In this chapter, we focus on the creation of a novel CAD framework that aims to automate the ABCD rules (Asymmetry, Border, Colour, and Dermoscopic structures) using SVM models and Bayesian fusion. To incorporate case-based reasoning, an Artificial Neural Network (ANN) is implemented to identify skin lesions with similar features.

This chapter proposes a CAD framework for the detection of melanoma using data extraction techniques to ensure the use of relevant features. The aim is to produce a transparent system focusing on providing information that would be useful to dermatologists and the impact of those features on the diagnosis. Metadata is included regarding age, gender, and anatomical location. Other features are asymmetry, border, colour, and dermoscopic structures. These are then combined using a Bayesian network. Case-based reasoning is also implemented to find skin lesions with similar clinical features.

7.2 Background

Automatic systems are being developed for the early detection of melanoma because it can take 10 years of experience for an accuracy of 86%[66]. Melanoma is one of the most aggressive forms of cancer that can remain dormant from anywhere between 6 months and 10 years before developing into metastatic melanoma, which becomes substantially more difficult to cure[103]. Problematically, clinicians who are not trained specifically to diagnose melanoma are usually the first to attempt it. Improving the accuracy of these clinicians should increase the overall accuracy of detecting melanoma. The early detection

of melanoma followed by a biopsy is known to completely cure the disease[65]. Furthermore, melanoma develops from melanocytes that create skin pigmentation through the production of melanin, making a brown patch on the skin. Therefore, it has a clear indication of development on the surface of the skin. This means it is ideal for the creation of computer vision models for early detection.

For these reasons, there has been further interest in developing an automatic system for helping clinicians detect melanoma at its early stages. However, regardless of newer systems being developed they are still rarely implemented within clinical environments. This is largely due to systems producing parallel diagnosis, which does not explain how results were reached[57, 10]. These techniques usually utilise Convolutional Neural Networks (CNN) because of their superior accuracy[108]. There should be further explanations of the diagnosis for clinicians to understand and properly utilise within clinical environments.

Newer machine learning models utilise explainable AI (XAI) to provide an explanation that provides further insight[92]. While these provide some indication of which area of the image has been used to train the algorithm they are still not tied directly to relevant clinical features. Furthermore, there appears to be a tradeoff between interpretability and model performance. Clinicians might not want to utilise models that are more interpretable, but less accurate. Furthermore, there has been some indication of models producing realistic but incorrect results[57]. In some scenarios, clinicians might be misled to falsely diagnose a skin lesion. Due to the high stakes involved when diagnosing melanoma, there should be highly accurate explanations and a track record of success before utilising them.

7.3 Related Work

In the design of CAD systems, two types of algorithms are employed. The initial type involves feature extraction, which is followed by individual classification methods. This process holds significant importance as it ensures the utilisation and visualisation of clinical features. The latter algorithm utilises the extracted features to classify different types of skin lesions. This method produces clinically relevant features that facilitate the diagnosis.

7.3.1 Feature Extraction algorithms

Ihab S. zaqout[113] describes a technique using the centroid and rotation of the skin lesion using moments of inertia. The skin lesion is folded over vertically and horizontally subtracting the opposite halves. Pixels that cannot be subtracted are summed and compared with a threshold. If over the threshold the skin lesion is considered asymmetrical in that direction.

Reda Kasmi and Karim Mokrani[46] describe a technique for comparing the colour distribution of the skin lesion by splitting the lesion into a 20 by 20 grid and comparing it against the colour of the perpendicular square using the 3D Euclidean distance. If over a

threshold that square is considered asymmetrical. If more than half the lesions are over the threshold then the area is considered asymmetrical.

7.3.2 Classification Methods

A paper described by Javier Lopez, et al.[58] describes a CAD system designed to provide clinicians with an enriched diagnosis. They utilise data extraction and classification techniques on dermoscopic structures, followed by combining the output of individual features using a Bayesian approach. This provides an indication of which features impact the diagnosis.

7.4 Proposed Method

The proposed CAD framework described in Figure7.1 automates the ABCD rules using statistical algorithms to extract features (f) from asymmetry, border, colour, and dermoscopic structures. Each feature has an associated SVM model trained using these extracted features. Next, Bayesian fusion, a probabilistic approach, combines multiple independent classifiers to diagnose melanoma. One benefit of Bayesian fusion is its higher accuracy in classifying skin lesions as compared to a standalone classifier[98]. Javier López-Labraca, et al describe a similar method using dermoscopic structures, and colour[58]. Other benefits are estimating the relevance of individual classifiers and classifying them with incomplete data, making it an interpretable and robust method. In addition, some feature extraction techniques generate graphics that might be suitable as an explanation for the diagnosis. Finally, the PH^2 dataset validates the rules, and once combined into a diagnosis, more extensive datasets, including ISIC 2019, measure its accuracy based on the diagnosis.

The ABCD rules, a set of criteria employed for the early detection of skin cancer, especially melanoma, are instrumental in guiding clinicians and individuals in assessing potentially suspicious skin lesions. The acronym "ABCD" stands for asymmetry, border irregularity, color variation, and diameter. Each element serves as a crucial parameter in evaluating the characteristics of moles or lesions on the skin. The objective of utilising this diagnostic procedure is to visualise important features that GPs need to support their diagnosis. Automating this process will instantiate trust when automating skin lesion identification within clinical environments. Another advantage would be the automatic labelling of skin lesions, making it easier for dermatologists to identify later.

The CAD framework in figure7.1 describes a model of pre-processing, feature extraction, and classification stages. After segmentation, statistical algorithms extract features (f), representing a different rule. Next, SVM models individually process the extracted features and combine them into a final result between benign and malignant using Bayesian fusion.

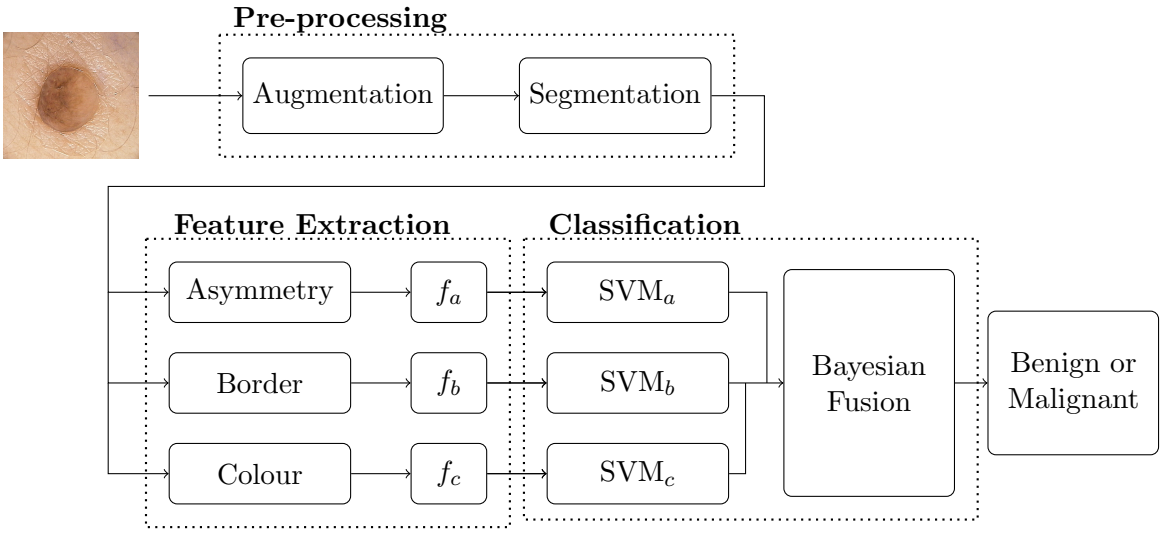


Figure 7.1: Proposed CAD framework describing the segmentation, feature extraction and classification process.

7.4.1 Feature Extraction Methods

ABCD rules in this refer to the asymmetry, border irregularity, colour variation, and dermoscopic structures. Sometimes the D in ABCD rules refers to the diameter of the skin lesion, but it is often removed for dermocopic structures because images are taken at different distances making the measurement of diameter unreliable. Furthermore, the detection of dermoscopic structures provides valuable information in detecting the melanoma mimic called seborrhoeic keratosis (SK)[62].

Asymmetry

The approach for identifying asymmetry in this chapter is adapted from the work of Reda Kasmi and Karim Mokrani[46]. In the original method, a bi-fold technique is employed to determine the centroid and rotation of the skin lesion, followed by image rotation. Subsequently, the image undergoes a conversion to LAB colorspace and is divided into a 20 by 20 grid by averaging color areas. The modified technique, however, utilizes superpixels from Simple Linear Iterative Clustering (SLIC) introduced by Achanta et al.[3], with a compactness (C) set to 20.

Each color square is compared with others perpendicular to the centroid (with bi-fold) using the 3D Euclidean distance of the color (LAB). The resulting distance score is accumulated in an array, and if more than half exceed a threshold of 6, the lesion is classified as asymmetrical, contributing a TDS score of 1. This process is repeated at a 90-degree orientation, adding another TDS score if asymmetry is detected.

The data presented in Figure?? illustrates that employing superpixels better supports the thresholding process compared to the original technique. The boxplot data represents the Euclidean distance, summed and divided by the number of positions.

Colour

Dermoscopic Structures

7.4.2 Bayesian Fusion using Naive Bayes

Bayesian fusion is a class of methods used to combine information from multiple sources taking into account uncertainty and probability distributions. This technique is frequently used for medical diagnosis for integrating data from various diagnostic tests to improve the accuracy of disease diagnosis[**empty citation**].

7.4.3 Case-Based reasoning using Artificial Neural Network (ANN)

7.5 Results

Two datasets were utilised to test the produced algorithms. The first is the PH2 dataset which includes asymmetry, colour, and dermoscopic structures.

7.6 Discussion

7.7 Conclusion

Chapter 8

Conclusion

Validating the automatic ABCD rules is challenging because public datasets are scarce and often lack sufficient data. For example, PH^2 contains 200 images on asymmetry, colour, and some dermoscopic structures but misses border irregularity. Therefore researchers aiming to measure borders use private or privately annotated datasets. Furthermore, many papers measuring asymmetry, colour and dermoscopic structures lack validation using public datasets despite PH^2 being available at the date of their publication. On the other hand, public datasets are crucial to comparing, validating, and reproducing algorithms. Therefore ABCD rules (apart from the border) will be validated using PH^2 datasets so that future researchers can replicate techniques. Furthermore, once rules are combined using Bayesian fusion, a type of probabilistic analysis, results can conform to the diagnosis between malignant and benign, validated from larger datasets, including ISIC 2019.

Finding the border cut-off is fundamental for the classification of melanoma using the ABCD rules[71]. Many valuable techniques use statistical models, including LBPC and Otsu, instead of transposed CNNs such as SegNet. Hybrid approaches using SegNet followed by Otsu to measure the border cut-off have been proven beneficial. However, using SegNet without a statistical model is worse when used with the ABCD rules than current methods such as LBPC and Otsu. Therefore, exploring other statistical segmentation techniques and hybrids would be beneficial. Furthermore, segmentation ground-truths do not always correspond to good classification accuracy with ABCD rules, which means even a low accuracy segmentation compared to datasets might have better accuracy when classifying the ABCD rules for border irregularity.

Statistical models for asymmetry, border, and colour extract relevant features for melanoma classification. The goal is to mimic the diagnostic procedure that clinicians are familiar with to produce results that they can utilise in a clinical environment. Extracting relevant features using box-counting and bi-folds ensures capturing relevant features and that the technique is retractable. However, accuracy is lacking in these techniques where superpixels improved asymmetry, changing the accuracy from 58.5% to 61% for the PH^2

dataset. Further improvements will be made after training an SVM model using the extracted features. Further implementation of convexity and Zernike moments for border irregularity will improve the accuracy. Furthermore, implementing a texture comparison for asymmetry measurements improve accuracy again.

Chapter 9

Future Work

Developing algorithms to extract features of ABCD rules is beneficial to GPs because it improves interpretability. Future work will involve extracting more features and training SVM models. For example, extracting more relevant asymmetry features will help classify asymmetry as there is currently no unification of shape, colour, and texture into a single classification model. The extracted features will be combined into a diagnosis between benign and malignant using a Bayesian probabilistic network. Bayesian probability is beneficial because its highly accurate[98] and modifiable and ability to classify with incomplete data. For example, asymmetry, border, and colour are sometimes enough to classify skin lesions. However, in some cases, dermoscopic structures or other meta-data, including age, gender, touch, feeling, and location on the body, are required for an accurate diagnosis. Furthermore, This might benefit GPs because it encourages considering a wide range of not always considered features.

Melanoma evolves from benign lesions at initially 30%-50%, and despite its significance, clinicians or computers are not yet able to reliably predict this change. AI trained on relevant images could predict melanoma before it occurs[93]. Data on skin lesion evolution is rare in public datasets. However, the associated organisation has taken images of the same skin lesion multiple times. It would be incredibly beneficial to assess the quality of these images, which could potentially lead to the development of a technique describing evolution. Considering evolution in machine learning techniques in the future would be incredibly beneficial to the early detection of melanoma but can only be achieved when there is more data.

Chapter 10

Tables

Chapter 11

Appendix