Artificial Intelligence for Skin Lesion Analysis based on Computer Vision and Deep Learning

Thesis submitted in accordance with the requirements of
the University of Liverpool for the degree of Doctor in Philosophy by

Saeed Alzahrani

February 2023

Abstract

Skin lesions appear in various sizes and forms and can be localised in one place or spread across the whole body due to different conditions. Dermatologists typically undertake physical examinations to diagnose skin lesions. However, this task costs time and requires excessive effort and can be inconsistent. Depending on the type of lesion and whether or not malignancy is present, additional diagnostic testing, such as imaging or biopsy, may be needed. Computer-aided diagnosis (CAD) systems, using clinical and dermoscopic images, could provide a quantitative assessment tool to help clinicians identify skin lesions and evaluate their severity. The recent progress in computer vision and deep learning has encouraged researchers to harness medical imaging data to develop powerful tools which could provide better diagnosis, treatment and prediction of skin conditions.

By leveraging artificial intelligence techniques, including computer vision and deep learning, this work introduces intelligent computerised approaches using dermoscopic and clinical images to analyse and identify two types of skin lesions producing enhanced medical information. This thesis designed, realised, and evaluated the benefit of features learned automatically from images through the stacked layers of convolution filters in the convolutional neural network (CNN) models. The final objective of conducting the research in this thesis is to benefit patients with skin lesion condition assessment and skin cancer identification without adding to the already high medical costs. An automated regression-based method has been developed in this thesis for acne counting and severity grading from clinical facial images. In addition to the acne lesions, another type of skin lesion has been considered, represented by melanoma-related lesions. Two pipelines have been presented in this thesis to identify melanoma lesions. The first framework benchmarks and evaluates several CNN models for melanoma and non-melanoma classification from only dermoscopic images. While the second developed model for melanoma detection integrates the seven-point checklist scheme with CNN using both clinical and dermoscopic images.

The experimental results of the work presented in this thesis manifest improved/-competitive performance compared to the state-of-the-art skin analysis methods using several evaluation metrics. The findings of the developed approaches demonstrated effective analysis of skin lesions with high accuracy, reducing the risk of misdiagnosis, and providing a more efficient means of detecting melanoma and automated acne lesion severity grading. Additionally, the application of computational intelligence allows for cost savings by reducing the need for manual analysis and enabling the automation of grading support, resulting in a more reliable and consistent process. Overall, the new automated methods based on computational intelligence demonstrate the benefits of developing computer vision and deep learning techniques for skin lesion analysis towards early skin cancer identification and cost-effective and robust grading support.

Contents

Illustrations

List of Figures

List of Tables

List of Abbreviations

The following abbreviations are found throughout this thesis:

ABCDE:	Asymmetric, Border, Shape, Color, Diameter, and Evolution
Acc:	Accuracy
AI:	Artificial Intelligence
ANN:	Artificial Neural Network
B:	Benign
BB:	Bounding Boxes
BCC:	Basal Cell Carcinoma
BCE:	Binary Crossentropy
BWV:	Blue Whitish Veil
CAD:	Computer-Aided Diagnosis
CCE:	Categorical Crossentropy
CNNs:	Convolutional Neural Networks
CUDA:	Compute Unified Device Architecture
D:	Detection
DBNs:	Deep Belief Networks
DEJ:	Dermal-epidermal Junction
DL:	Deep Learning
Err:	Error
FC:	Fully Connected Layers
FN:	False Negative
FNR:	False Negative Rate
FP:	False Positive
FPR:	False Positive Rate
GANs:	Generative Adversarial Networks
GPU:	Graphical Processing Units
ISIC:	International Skin Imaging Collaboration
LD:	Label Distribution
LSTMs:	Long Short Term Memory Networks

M:	Malignant
MCC:	Mathews Correlation Coefficient
MCDM:	Multi-Criteria Decision Making
MLPs:	Multi-layer Perceptrons
MSE:	Mean Squared Error
NLP:	Natural Language Processing
NP:	Not Applicable
NPD:	Non-polarised Dermoscopy
OCT:	Optical Coherence Tomography
OVO:	One-vs-One
OVR:	One-vs-Rest
PD:	Polarised Dermoscopy
PIG:	Pigmentation
PROMETHEE:	Preference Ranking Organization Method for Enrichment of Evaluations
R-DL:	Regression based Deep Learning
R-ML:	Regression based Machine Learning
PN:	Pigment Network
Pre:	Precision
RBMs:	Restricted Boltzmann Machines
ReLu:	Rectified Linear Unit
RNNs:	Recurrent Neural Networks
RPN:	Region Proposal Network
RS:	Regression Structures
SCC:	Squamous Cell Carcinoma
SGD:	Stochastic Gradient Descent
Sn:	Sensitivity
Sp:	Specificity
STR:	Streaks
SVM:	Support Vector Machine
TN:	True Negative
TP:	True Positive
VIKOR:	VIseKriterijumska Optimizacija I Kompromisno Resenje
VS:	Vascular Structure

Acknowledgements

When I think about the past four years, I would say that pursuing a PhD has been tough and uncertain, but I have ended up finding a way to cope with it and accomplish my thesis. This way was a difficult quest to find. I deeply appreciate everyone who helped and supported me through this experience.

I would like to start by sincerely thanking my supervisor, Dr Waleed Al-Nuaimy, for his unconditional support of my PhD study and his patience, inspiration, and deep understanding and knowledge. His guidance and advice were incredibly helpful to me during the entire research and thesis-writing phases. I could not have wished for a better supervisor to help me with my doctoral work.

Additionally, I would like to express gratitude to my colleagues, Dr Baidaa Al-Bander and Dr Theiab Alzahrani, for their treasured support, which influenced my experiment methods and critiquing my results and thesis writing up. My gratitude extends to the Kingdom of Saudi Arabia government for the funding opportunity to undertake my studies at the Department of Electrical Engineering and Electronics, University of Liverpool.

I would like to express my gratitude to my parents, wife, and children for supporting me through tough and challenging circumstances. Without their wonderful support and understanding throughout the previous few years, I would not have been able to finish my research work.

Chapter 1

Introduction

1.1 Background and Motivation

Decision-makers in the healthcare sector are constantly subjected to extreme pressure to provide high-quality treatment with constrained capacity and limited resources. Healthcare providers have a significant difficulty due to the shortage of trained workers. Nowadays, cutting-edge technologies in healthcare can provide new means for enhancing workers' efficiency and patient diagnosis outcomes. The steadily-increasing digitisation of healthcare and linking it with data has encouraged researchers to harness the data-driven methods and potentials of artificial intelligence (AI) for the healthcare industry. By exploiting data acquired and generated in daily clinical practice, AI-supported services could support clinicians in assessing and analysing many medical and clinical conditions, including skin lesions. A skin lesion can be defined as the abnormal appearance of the skin in comparison to the surrounding skin [1, 2].

Skin imaging tools are a very crucial part of skin-related research. To date, many popular skin imaging modalities have been developed and designed. Dermatoscopy is one of the most common examination methods adopted to assess, and screen skin lesions by a dermoscopy modality [3, 4]. Recently, several datasets have become publicly accessible to aid the study and development of automated skin image analysis using a variety of imaging modalities, including dermoscopic and clinical photographs. These advancements have sparked a surge of interest in skin image analysis research. With the

advancement of skin imaging technologies and AI techniques, especially computer vision and deep learning represented by conventional neural networks (CNNs), there is a genuine need for developing computer-aided diagnosis (CAD) systems to help towards skin lesion analysis. These automated approaches could provide dermatologists with means and intelligent assisted schemes to identify skin lesions from dermoscopy and clinical photographs. The automatic skin lesion analysis methods and CAD systems could significantly reduce the workload of manual examination and grading and thus help towards decreasing undesirable costly screening.

1.2 Problem Statement

This thesis deals with two types of skin lesions: facial acne-related lesions and melanoma-related lesions. The research problem of these two lesion types, both clinically and technically, is stated as follows.

Facial acne vulgaris is the common form of acne that primarily affects adolescents and might persist into adulthood, characterised by an eruption of inflammatory and/or non-inflammatory skin lesions. The psychosocial consequences of acne scars can be exceedingly detrimental, and they may be a risk factor for major mental issues [5, 6]. The effectiveness of acne treatment is usually established by the physician's comprehensive and valid assessment. For assessment by a physician, different forms of acne lesions need to be counted and examined independently. On the other hand, manual acne lesion evaluation can be difficult and time-consuming, considering the limited consultation time. Autonomous acne identification, counting, and evaluation systems would help dermatologists achieve a more reliable and consistent examination of acne in clinics, thanks to the advancement of deep learning, computer vision, imaging technology, and widespread access to cameras [7].

Generally speaking, developing automated methods for acne detection and severity grading is significant because it offers several benefits over manual analysis. Firstly, it is a more efficient and quicker process, reducing the time required for diagnosis and

treatment. Secondly, automated methods have the potential to be more accurate and consistent, reducing the risk of misdiagnosis and ensuring that patients receive the appropriate treatment. Additionally, computerised techniques can help reduce the costs associated with manual analysis, making it a more cost-effective solution for healthcare providers. Furthermore, these methods can help standardise the grading process, providing a more objective and reliable assessment of acne severity. Ultimately, the development of automated acne detection and severity grading methods can significantly improve the quality of care for patients and help healthcare providers provide more effective and efficient treatments.

The most common kind of cancer is skin cancer, which can be extremely malignant. Melanoma is the most dangerous type of skin cancer. It grows quickly and has the ability to spread to any organ. While it is serious skin cancer, it is highly curable if detected early. Melanoma diagnosis is difficult, even for experienced dermatologists, due to the wide range of morphologies in skin lesions [8]. The process of automatically recognising melanoma lesions from dermoscopy photographs is challenging due to various complexities. First, it is difficult to precisely split lesion areas due to the poor contrast between healthy skin and skin lesions. Secondly, benign and malignant lesions sometimes have a significant visual similarity, causing both lesions challenging to be identified. Thirdly, individuals with diverse skin traits, such as natural skin colour, have different melanoma-related lesions appearance in terms of texture and colour [9].

1.3 Aim and Objectives

The main aim of the research work conducted in this thesis is to develop robust, reliable, and intelligent computerised approaches based on computer vision and deep learning algorithms for skin lesion analysis. Developing such automated lesion analysis techniques could help physicians in making their decisions regarding diagnosis and reduce examination time. Achieving the research aim requires designing various advanced image analysis tasks, including object detection, regression, and classification using deep

learning algorithms. These techniques have been exploited through the thesis to help develop automatic feature extraction methods from skin lesion images for skin condition and cancer identification. The research themes, considering the analysis of two types of skin lesions, including acne- and melanoma-related lesions, are presented as follows:

1. **Grading of Skin Lesions Related to Acne From Facial Images using Regression-based Deep Learning Model:** Facial acne lesion counting and severity grading using regression-based deep learning model.

2. **Classification of Skin Lesions Related to Melanoma From Dermoscopic Images:** Benchmarking and evaluation of CNN classification architectures for melanoma lesion detection.

3. **Classification of Skin Lesions Related to Melanoma From Dermoscopic and Clinical Images via Seven-point Checklist Criteria:** Integrating Seven-point checklist with CNN classification architecture for melanoma lesion detection.

1.4 Hardware and Software

The developed systems in this thesis have been built up and designed using a variety of deep learning frameworks and computer vision libraries. For effective convolutional neural network development and training, these frameworks reduce the time-consuming procedure of starting learning from scratch and allow parallel computation in graphical processing units (GPU). The compute unified device architecture (CUDA) programming model, made available by NVIDIA[1], allows for heavy parallel computations. These heavy computations are typically required by deep learning models. In this thesis, the experimental work was carried out on a workstation (HP Z440) running the Linux operating system and equipped with a 3.50GHz Xeon CPU, 12GB GTX TITAN

[1] `https://www.nvidia.com/en-me/geforce/`

X GPU, and 16GB RAM. Tensorflow[2], OpenCV[3], Scikit-Learn[4], NumPy[5], SciPy[6], and Matplotlib[7] are python libraries which are widely used by researchers. In chapters three and five of this thesis, these libraries were used to implement image processing methods, deep learning models, and visualisation strategies. Methods developed in chapter four have been implemented under the MATLAB environment using the same hardware and GPU specifications used in the rest chapters. Statistics and Machine Learning, Deep Learning, Image Processing and Computer Vision are the MATLAB toolboxes imported to develop the work in chapter four.

1.5 Summary of Contributions

In general, the contribution of the presented work can be described as follows:

- An effective automated attention mechanism integrated with dilated UNet regressor for acne counting and severity grading from two-dimensional facial images. The main contributions of this method can be described as follows:
  - Inspired by the scenario of crowed counting from kernel density maps and leveraging the advances of deep learning models, a new method for acne counting and severity grading called dilated UNet dense regressor guided by attention mechanism was developed.
  - Modifying the paths of contraction-expanding (encoder–decoder paths) in the UNet segmentation model by introducing a bounding box encoder that incorporates the box information generated by Faster R-CNN.
  - This embedding adaptation helps to simultaneously handle high- and low-density regions of acne lesions.

[2] https://www.tensorflow.org/
[3] https://opencv.org/
[4] https://scikit-learn.org/stable/
[5] https://numpy.org/
[6] https://scipy.org/
[7] https://matplotlib.org/

  - The proposed regressor exploits dilated convolutions to aggregate multi-scale contextual details systematically.
  - Experiments on public facial acne image datasets demonstrate the superiority of the proposed method compared with the state-of-the-art techniques.

- Conducting a comprehensive evaluation and benchmark of convolutional neural networks for melanoma diagnosis. The contribution of this work can be represented in three-fold:

  - The proposed study provides an appropriate and powerful linkage between the multi-criteria decision-making techniques and the objective performance evaluation criteria, which are typically used to evaluate the deep learning models. This integration with decision-making schemes helps rank the learning models based on multiple conflicting criteria and select the optimal model in the presented case study.
  - This is the first study that introduces the application of a multi-criteria decision-making approach based on merging entropy and PROMETHEE methods to help prioritise the deep convolutional neural networks used for melanoma diagnosis and select the optimal model considering various criteria.
  - This study presents a comprehensive evaluation of nineteen convolutional neural network models with a two-class classifier. The models are trained and evaluated on a dataset of 991 dermoscopic images considering ten performance evaluation metrics.

- A deep learning-based method has been proposed to predict the 7-point checklist criteria [10] and diagnose melanoma where the lesion features are designed automatically. Multiple input convolutional neural networks (CNNs) considering clinical and dermoscopic images as inputs have been developed. The incorporation of 7-point checklist criteria with CNN as well as learning the proposed model

using difficult and non-standardised images (clinical images), may aid with leveraging the reliability of melanoma diagnosis.

1.6 Publications

The results of the research work conducted in this thesis have been published in two peer-reviewed conferences and two journals, as follows.

1. **Saeed Alzahrani**, Waleed Al-Nuaimy, Baidaa Al-Bander, "Seven-Points Checklist with Convolutional Neural Networks for Melanoma Diagnosis ", in 2019 8^{th} *European Workshop on Visual Information Processing (EUVIP). IEEE, 2019, pp. 211–216*.

2. **Saeed Alzahrani**, Waleed Al-Nuaimy, "Deep Learning Approach for Skin Lesion Attributes Detection and Melanoma Diagnosis ", in 2020 2^{nd} *International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI). IFSA, 2020, pp. 222-223*.

3. **Saeed Alzahrani**, Baidaa Al-Bander, Waleed Al-Nuaimy, "A Comprehensive Evaluation and Benchmarking of Convolutional Neural Networks for Melanoma Diagnosis ", Cancers 2021, 13, 4494. https://doi.org/10.3390/cancers13174494.

4. **Saeed Alzahrani**, Baidaa Al-Bander, Waleed Al-Nuaimy, "Attention Mechanism Guided Deep Regression Model for Acne Severity Grading ", Computers 2022, 11, 31. https://doi.org/10.3390/computers11030031.

1.7 Structure of the Thesis

The organisation of this thesis can be described as follows:

Chapter 2 covers an overview of the theoretical background of deep learning models and relevant clinical and technical skin lesion analysis methods and terminologies.

Chapter 3 describes the proposed attention-guided UNet dense regressor for addressing the task of acne counting and severity grading in detail. The developed architecture incorporates dilated UNet dense regressor for density regression with the information of bounding boxes generated from Faster R-CNN network, producing a hybrid detection–regression framework.

Chapter 4 presents the developed evaluation and the benchmarking system, which comprises five main stages, including data preparation, designing of CNN models, training of CNN models, evaluation criteria establishment, and benchmarking of CNN models using Multiple Criteria Decision Making (MCDM).

Chapter 5 proposes a new technique for skin lesion detection and melanoma diagnosis from dermoscopy and clinical images by combining seven-point checklist criteria with convolutional neural networks.

Chapter 6 discusses and concludes the research work presented in this thesis, and provides suggestions for potential future research themes.

Chapter 2

Literature Review

2.1 Skin Lesions

The biggest organ in the body is the skin. It offers defence against diseases, light, heat, and cold. The epidermis and dermis, two primary skin layers and other cell types, make up the skin. The epidermis, which is the top layer of skin, is made up of three kinds of cells: melanocytes, which give skin its colour and defend against abrasion; round cells called basal cells; and flat, rough surface cells called squamous cells. The nerves, blood vessels, and sweat glands are located in the dermis, the skin's innermost layer [11]. The area of medicine that deals with both medical and surgical aspects of skin is called dermatology. Figure 2.1 depicts the skin structure and layers.

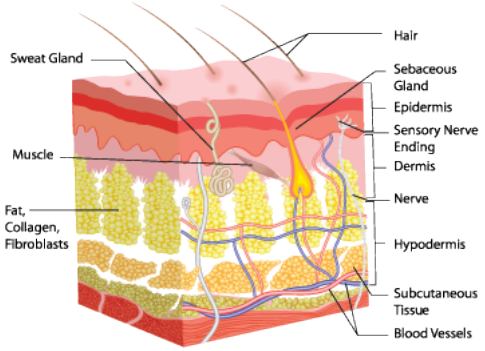


FIGURE 2.1: The layers of human skin [11].

A skin lesion is any area of skin that differs in size, colour, shape, or texture from the surrounding area of skin. Skin lesions are relatively prevalent and frequently develop from a localised skin injury, such as contact dermatitis or sunburn. While others, including diabetes, infections, and autoimmune or genetic illnesses, might be symptoms of underlying conditions. Most skin lesions are benign and painless, but some have the potential to become malignant or pre-malignant, which means they might turn into skin cancer [12].

2.1.1 Primary Skin Lesion Types and Causes

There are two basic types of skin lesions, primary and secondary lesions. The first type, primary skin lesion, develops on initially healthy skin and has a known aetiology. Freckles, acne, rashes, nodules, pustules, moles, and blisters are a few examples of primary skin lesions that are frequently seen [13]. On the contrary, secondary skin lesions emerge from a primary skin lesion as it progresses or as a result of traumatising manipulation, such as scratching or rubbing. Crusts, sores, skin atrophy, ulcers, and scars are a few examples of secondary skin lesions [12, 14].

Many conditions can cause different types of skin lesions. Primary skin lesions can be seen as tumours or non-tumours (for instance, acne lesions). Common skin tumour growths can either be non-cancerous, also called benign (typically very slow growing and without spreading to other places) or cancerous, also called malignant (generally growing extremely fast and potentially invading surrounding tissues). Both types are described as follows:

1. **Tumours**

    - Benign Skin Lesions

      An irregularity, growth, or tumour of the skin that is not malignant is referred to as a benign skin lesion. Depending on its aetiology, benign lesions can present in a variety of ways. Most melanocytic nevi, also known as

moles, skin tags, seborrheic keratoses, lipomas, and cherry angiomas, are examples of common benign skin lesions. When these lesions do not create symptoms like discomfort or itching, they are often not harmful and do not need treatment. It is typical for benign skin lesions to be symmetrical, well-circumscribed, consistent in appearance, stable, or slowly growing. Although it can be challenging to discern between benign and malignant lesions in some situations, in those contexts, a biopsy or surgical excision of the afflicted region can be done to verify malignancy. [12]

Additionally, premalignant lesions like actinic keratosis and lentigo maligna, which have a higher chance of turning into various forms of skin cancer, should be differentiated from benign lesions. Lentigo maligna and actinic keratosis are brought on by prolonged, excessive sun exposure. While lentigo maligna shows as localised dark-brown or black lesions, often on the face and trunk, actinic keratosis manifests as dry, scaly patches of skin across sun-exposed regions, such as the nose and forehead [12].

- Malignant Skin Lesions

  A malignant skin lesion is, by definition, skin cancer. Melanoma and keratinocyte carcinoma (non-melanoma) are the two primary kinds of skin cancer. While each kind of skin cancer has its own features, typical indications include lesions on the skin that are developing quickly, changes in the colour or size of an existing lesion, or scabbing sores that do not heal over time. Basal cell carcinoma (BCC) and squamous cell carcinoma (SCC) are two kinds of keratinocyte carcinoma that develop from skin cells called keratinocytes. A skin lesion with pearly, flesh-coloured skin and telangiectasias—superficial blood vessels— may seem to be basal cell carcinoma. Basal cell carcinoma might appear as a crusted or bleeding non-healing sore or as a superficial scaling plaque. Squamous cell carcinoma, on the other hand, typically presents as a thick, crusty sore with a reddish, inflammatory

base that can ulcer (look like an open sore) and bleeds. Melanocytes, which are skin cells, lead up to melanoma. Melanoma often has an unusual or uneven appearance [12].

Skin cancer is the most frequent type of cancer and can be highly truculent [8]. In the UK, more than 100,000 new cases of skin cancer are reported each year [15]. In 2016, 1319 death cases from non-melanoma skin cancer and 2285 death cases from melanoma skin cancer were reported [16, 17]. Non-melanoma skin cancer, including squamous cell carcinoma (SCC) and basal cell carcinoma (BCC), are the vast majority of skin cancers. These are unlikely to spread to the remaining parts of the human body; however, they may be locally disfiguring if not diagnosed and treated early. In contrast, malignant melanoma is a minor skin cancer type but a fetal and highly aggressive, which tends to spread to the other parts of the body, causing death if it is not diagnosed and treated early [18, 19].

2. **Acne Vulgaris**

Acne vulgaris, or acne, is a skin condition in which dead skin cells and oil from the skin block hair follicles. This skin condition is clinically featured by blackheads and whiteheads (open and closed comedones), small and tender red bumps (papules), white or yellow squeezable spots (pustules), cyst-like fluctuant swellings (cysts), and large painful red lumps (nodules), as shown in Figure 2.2. It usually affects areas of skin with a high number of oil glands, such as the face, chest, back and shoulders [20, 21]. Facial acne is most common during adolescence, but it can persist into adulthood. After severe inflammatory acne, scarring inevitably occurs. The scarring might lead to significant psycho-social consequences and potential risk factors for serious mental health issues. The resultant facial appearance can cause anxiety, low self-esteem, and, in the worst-case scenario, depression or suicidal thoughts [5, 6]. Treatments include medications, and sometimes laser or light therapy [22].

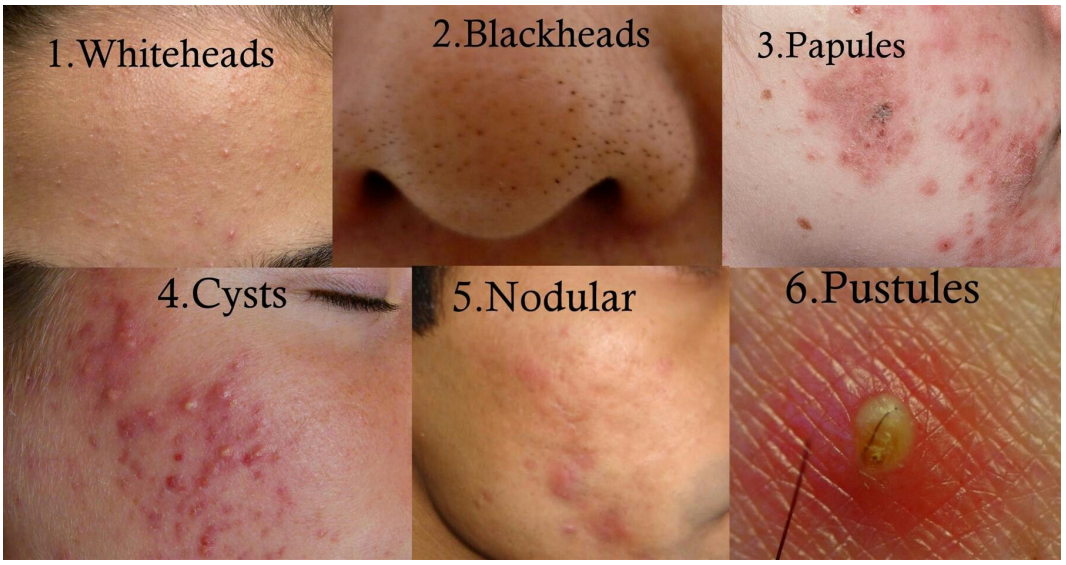


FIGURE 2.2: Image example shows acne lesion types [23].

2.2 Imaging in Dermatology: Skin Imaging Modalities

Imaging technology has been used extensively in medicine ever since x-rays were first developed. Different imaging modalities have been created as a result of technological advances, the majority of which have been utilised to examine organs located deep inside the body. Recently, the use of imaging technologies for skin examination has attracted a lot of interest. Specialised photography, ultrasound, surface microscopy, optical imaging, confocal microscopy, laser Doppler perfusion imaging, and magnetic resonance imaging are among the methods now being utilised to investigate the skin. These techniques can give insights that can help in the treatment of skin issues [24].

Optical imaging, a non-invasive imaging technology for skin cancer detection, is one of the newest medical technologies. The use of technology to examine bodily structures using visible light and photons is known as optical imaging. Because optical imaging employs non-ionizing radiation, the patient is exposed to considerably less radiation. Due to the non-invasive nature of the technology, there is no need for biopsies or surgical incisions to obtain the outcome of the diagnosis. This technique could distinguish between soft tissues that are likely to be cancerous and those that are healthy. In addition to helping identify various skin disorders and lesions, these new approaches could

be very useful for the early diagnosis of skin cancer [25, 26]. Several distinct optical imaging methods are now being employed or explored for their potential application in the detection of skin cancer [25, 27], including:

- Dermoscopy: it produces microscopic images in a realistic scenario that can identify an abnormality in the skin's superficial dermis and epidermis layers using polarised light and a high-quality magnifying lens. Dermoscopy has been shown to be useful in diagnosing many lesions, including squamous cell carcinoma, basal cell carcinoma and actinic keratosis.

- Optical coherence tomography (OCT): real-time scans of the structures below the skin are produced using optical coherence tomography (OCT) technology. Squamous cell carcinoma, basal cell carcinoma and actinic keratosis can all be diagnosed using optical coherence tomography (OCT), which has been utilised in ophthalmology since 1991.

- Florescence photography: doctors can use cross-polarised light and florescence photography to detect lesions like actinic keratosis and basal cell carcinoma that can be invisible to the naked human eye.

- Confocal microscopy: produces a real view of the intra-cellular components of several skin layers using a strong microscope and light from a laser. The confocal microscope, despite its high cost, is becoming increasingly popular because of its capacity to identify skin conditions non-invasively. Confocal microscopy is particularly helpful in identifying squamous cell carcinoma, basal cell carcinoma and actinic keratosis, as well as distinguishing between atypical moles (nevi) and melanoma.

- High-frequency ultrasound: for skin lesion identification purposes, including benign tumours, high-frequency ultrasound utilises ultrasonic signals which pass across the skin layers and reflect an image, enabling clinicians to examine the various layers and underlying structures.

2.3 Skin Condition Diagnosis

2.3.1 Melanoma and Non-Melanoma Diagnosis

The initial stage in melanoma diagnosis is usually a visual assessment of the skin lesions. In comparison to inspection with the naked eye, dermatoscopy is one of the dermatologists' most popular imaging procedures, and a frequently used diagnostic tool that enhances and improves the diagnosis of malignant and benign pigmented skin lesions [28]. Contrary to dermoscopy images, clinical images are produced by capturing a direct snapshot of the skin disease location using a camera. They can act as an additional patient's medical record and offer various perspectives on dermoscopy images [29]. Because of the impact of various imaging conditions (i.e. illumination and angle of capturing), clinical images used for skin cancer identification have some drawbacks of providing poor morphological information while simultaneously bringing flaws into the diagnostic conclusions [30]. Figure 2.3 shows an example of a clinical image versus a dermoscopic image.

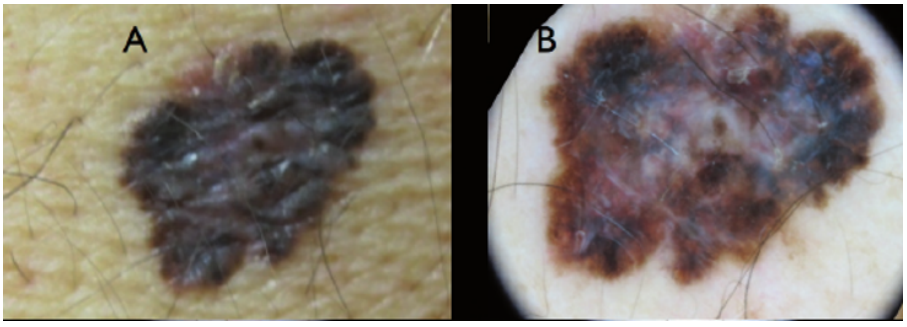


FIGURE 2.3: Clinical image (left) versus dermoscopy image (right). Adapted from [31].

A dermoscopy magnifies the surface of the skin lesion, allowing better visualisation of deeper skin structures. It provides improved diagnostic accuracy of skin lesions, enabling the dermatologist to examine them more thoroughly. There are two main dermoscopy modes: non-polarised dermoscopy (NPD) and polarised dermoscopy (PD).

Non-polarised dermoscopy (NPD) is integrated with a magnification lens and light-emitting diodes to provide illumination, enabling the visualisation of subsurface structures in the epidermis. Non-polarised dermoscopy (NPDs) require direct contact of the glass plate with the skin surface and the presence of a liquid interface such as alcohol, liquid paraffin, water, or ultrasound gel. Interface fluid dramatically increases the penetration of light, reduces scattered radiation, and produces a clear, low-reflection image which allows excellent visualisation of the superficial layers of the skin from the epidermis to the dermal-epidermal junction (DEJ). Like NPD, Polarised dermoscopy (PD) contains light-emitting diodes to provide illumination and are equipped with a magnification lens. However, PDs use two polarised filters to achieve cross-polarisation. NPD does not require direct contact with the skin and does not require the use of immersion liquids. PD allows the visualisation of subsurface structures located at the dermal-epidermal junction (DEJ) or superficial dermis. PD nearly blinds to the skin's surface and structures in the superficial epidermis. Hence, non-polarised dermoscopy reveals superficial features while polarised dermoscopy shows deeper structures, inferring that the use of both methods can provide complementary information [32, 33].

Melanoma is diagnosed in two ways: visual inspection and biopsy. ABCDE (Asymmetric, Shape, Border, Color, Diameter, and Evolution) [34] are the main criteria used for the visual screening of melanoma lesions based on a geometric description. Because the ABCDE approach is entirely dependent on the practitioner's visual acuity and experience, this approach can be performed efficiently only by trained dermatologists [35]. Seven-point checklist method [10] is also one of the most commonly recommended and accepted skin cancer visual-based assessment strategies [36]. Seven-point checklist method was established by Argenziano et al. [10] for the dermoscopic differentiation between benign and malignant lesions.

2.3.2 Acne Diagnosis and Severity Assessment

Acne vulgaris is simple to diagnose; however, its polymorphic structure makes it difficult to assess its severity. As the number of acne lesions varies during the course of the condition, numerous evaluation criteria based on clinical screening and photographic documentation have been established. Grading based on clinical examination, lesion counting, and approaches requiring instruments, such as photography, fluorescent photography, polarised light photography, video microscopy, and sebum production measurement, are developed to assess the severity of acne vulgaris. Clinical examination (grading) and lesion counting are two widely used methods for acne severity assessment [21, 37]. Clinical grading is a subjective approach that entails analysing the dominating lesions, assessing the occurrences of inflammation, and measuring the degree of involvement to determine the severity of acne. On the other hand, the acne lesion counting-based method involves counting the number of a certain kind of acne lesion and then evaluating the overall severity [37].

Acne severity has also been measured via photography, which involves comparing patients to a photographic standard. This method has many disadvantages, including the inability to palpate the depth of involvement and the difficulty of visualising small lesions. When it comes to determining the density of comedones, fluorescence and polarised light photography can offer some advantages over standard photography. However, there are some shortcomings, such as a substantial time commitment and the necessity for more complicated types of equipment [38]. In 2008, Hayashi et al. [39] presented a grading method to classify acne lesions into four types using standard photographs and lesion counting. On half of each patient's face, they counted the number of open and closed comedones, papules, pustules, cysts, and nodules. They categorised the eruptions into three groups: comedones, inflammatory eruptions (including papules and pustules), and severe eruptions (including cysts and nodules). They graded the severity of acne as (i) mild when the acne count is (0–5), (ii) moderate when the acne count is (6–20), (iii) severe when the acne count is (21–50), and (iv) very severe when the

acne count is more than 50, based on the number of inflammatory eruptions (papules, pustules) or lesions on half of the face.

2.4 Artificial Intelligence: Computer Vision and Deep Learning

The study of computer science that emphasises developing algorithms for carrying out activities typically thought to need human cognition and intellect is known as artificial intelligence (AI). Through the usage of intelligent technology, the industry has been developing and integrating into daily life. Numerous AI research fields, such as computer vision, machine learning and deep learning, natural language processing (NLP), robotics, expert systems, and fuzzy logic, have had practical applications for various scientific fields [40].

2.4.1 Computer Vision

In order to comprehend the concept of computer vision, it is imperative first to examine the definition of machine vision. Machine vision refers to the capability of machines, typically computers, to interpret and understand visual information in the same way humans do. It involves the use of cameras, algorithms, and software to process images or videos and extract meaningful information from them. Robotics, a branch of AI, is an example of a field that uses machine vision widely. An example of a robot that was designed based on machine vision systems is Sophia [41], one of the world's most well-known robots due to its advanced artificial intelligence and human-like appearance. Sophia has a realistic face with expressive eyes, eyebrows, and a mouth that can move and change expression, making it capable of mimicking human gestures and emotions. A camera is typically used in a machine vision system to capture images. Computer vision, the field of study within computer science and artificial intelligence and the software part of the vision system, then analyses and interpret the images by

applying certain algorithms before guiding other system components to respond to the provided information. It aims to develop algorithms and technologies that enable computers to recognise and understand visual content, including images, videos, and 3D scenes [42, 43].

Self-driving cars, which employ several cameras, radar, lidar, and sensors to scan the visual field around the vehicle and make driving judgments, are a common use of machine and computer vision. The use of machine and computer vision in healthcare is another significant application. Machine and computer vision helps healthcare professionals to detect disease symptoms significantly earlier and develops preventative strategies for maximising medical service outcomes. Applications of machine and computer vision in healthcare range from image analysis to surgical intervention and guidance. Automatic image processing, including medical images, is an interdisciplinary area that discusses how computers can interpret visual images or videos. Popular tasks involving the study of images involve classification, detection and segmentation. The algorithm aims at classifying images into two or more groups within the classification function. The algorithm seeks to localise structures in 2D or 3D image space within the detection task. In the segmentation function, the algorithm attempts to give an organ a pixel-wise delineation [42, 43].

2.4.2 Machine Learning and Deep Learning

Recent advances in science have made Machine Learning (ML) and Deep Learning (DL) well-known terms in artificial intelligence. An AI subclass called machine learning enables computers to learn without explicit programming. In conventional machine learning models, statistical approaches are employed to identify the category (class) information based on attributes extracted from data. Human scientists typically seek to find and extract the image features that best reflect the visual data.

In general, there are four categories of deep learning and machine learning algorithms: reinforcement learning, unsupervised learning, semi-supervised learning, and

supervised learning strategies. In the supervised learning scenario, data professionals provide algorithms with labelled training data as well as parameters to compute the correlations and discrimination in data samples. The algorithm's input and output used for training are pre-defined. In an unsupervised learning scenario, algorithms that learn from unlabelled data are built to infer or correlate significant correlations among data samples. Clustering algorithms using exploratory data analysis techniques are applied to find hidden patterns or groupings in datasets. In semi-supervised learning, semi-labelled datasets serve as the foundation for the learning process. This method has the benefit of utilising the least amount of labelled data available. In addition, this approach has certain drawbacks, including the potential for inaccurate outcomes due to irrelevant input features in training data. For the purpose of training a computer programme to complete a multiple-step procedure for which there are rules, a reinforcement learning scheme is used. Here, the learning model is designed to run a specific task and provide feedback (positive or negative) so that the algorithm responds to accomplish the task. The algorithm can decide what to do next to approach the final goal of the assigned job [44].

The way data is handled, analysed, and modified has altered due to recent developments in ML, particularly the newly developing topic of deep learning. The enthusiasm around deep learning, a cutting-edge field of artificial intelligence and a sub-type of machine learning results from the most representative and discriminative data attributes being hierarchically learned in an end-to-end fashion [45]. In deep learning techniques, a type of representation learning, no manually extracted feature set is used. The deep learning algorithm discovers on its own which attributes are most effective in categorising the data. Representation learning schemes in deep learning algorithms may outperform manually created features (also called hand-designed or hand-crafted features), providing sufficient training data. Figure 2.4 shows how the deep learning concept differs from machine learning and how they work. Deep learning techniques have greatly

succeeded in several real-world applications, including target detection and identification and robotics [46]. These techniques have significantly outperformed state-of-the-art methods in various disciplines and have gained significant attention from both the academic and industrial worlds [47]. The most popular deep learning models are Convolutional Neural Networks (CNNs) [48], Recurrent Neural Networks (RNNs) and Long Short Term Memory Networks (LSTMs) [49], Generative Adversarial Networks (GANs) [50], Deep Belief Networks (DBNs) [51], Restricted Boltzmann Machines (RBMs) [52], and Autoencoders [53].

Convolutional Neural Networks (CNNs) algorithm [48] is the most common deep learning model applied to solve several computer vision tasks, including classification [54], detection and localisation [55], segmentation [56]. They are well suited for image classification and recognition tasks due to their ability to capture local and spatial features. Recurrent Neural Networks (RNNs) [49] are good at handling sequential data such as time-series and speech recognition. However, they tend to struggle with capturing long-term dependencies. Long Short-Term Memory Networks (LSTMs) [49] address this limitation of RNNs and are commonly used in speech and language processing. Generative Adversarial Networks (GANs) [50] are used for generating new data similar to existing data. They consist of two parts: a generator and a discriminator. The generator generates data and the discriminator evaluates it. GANs can be used for image generation and super-resolution. Deep Belief Networks (DBNs) [51] are generative models that learn a probability distribution over the input data. They are mainly used for dimensionality reduction and feature learning. Restricted Boltzmann Machines (RBMs) [52] are shallow, two-layer generative models used for feature learning and dimensionality reduction. Autoencoders [53] are neural networks used for unsupervised learning and dimensionality reduction. They learn a compressed representation of the input data. CNNs have proven to be effective in medical image analysis and have shown promising results in identifying lesions, particularly in identifying abnormalities and detecting diseases [43, 57, 58]. Consequently, in the present work, it has been elected to concentrate on utilising Convolutional Neural Networks (CNNs) to

analyse medical image data. The concept and components of the CNNs are explored in the next section.

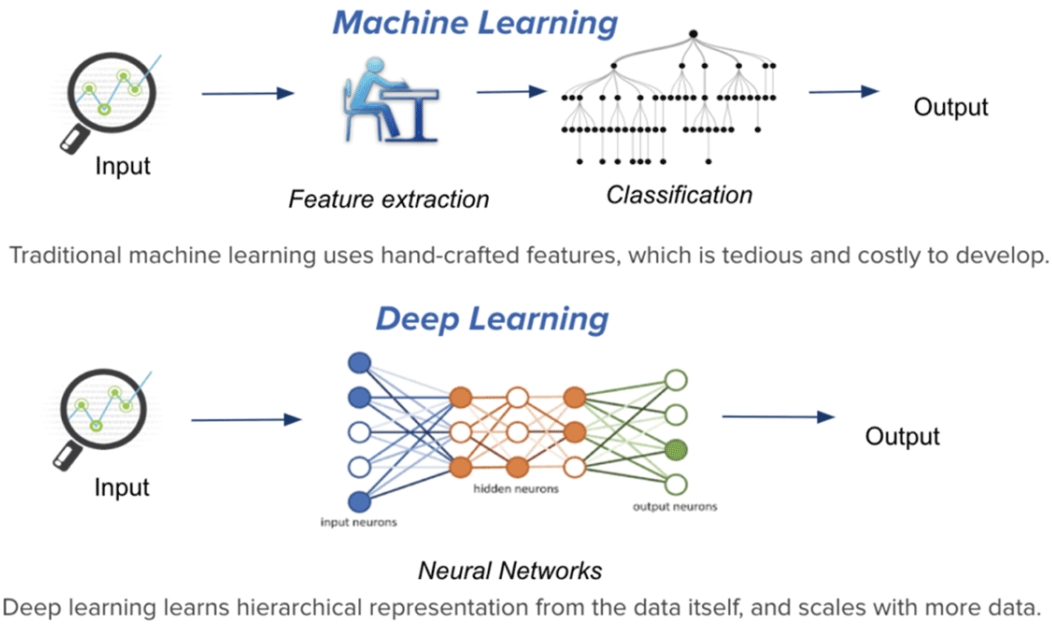


FIGURE 2.4: Machine learning and deep learning concepts [59].

2.4.3 Convolutional Neural Networks

Artificial convolutional neural networks (CNNs) are the most popular, well-established and often utilised technique for carrying out computer vision tasks in the domain of deep learning (DL), according to [60]. The fundamental advantage of CNN over its forerunners, such as artificial neural networks (ANNs), is the identification of the pertinent attributes automatically without human intervention. They have been widely used for a variety of industries and applications, including audio and speech processing [61] and computer vision [62, 63]. They are composed of a structure resembling traditional artificial neural networks, which were modelled to mimic the brain neurons of both animals and humans. Particularly, the CNNs simulate the complicated cell pattern that creates the visual cortex in a cat's brain [64]. According to Goodfellow et al. [65], "parameter sharing," "sparse interactions," and "equivalent representations" are the three main advantages of CNN. Contrary to fully connected networks (FC) that handle 1D input data, CNNs exploit the local connections and shared weights from 2D data structures, such as

image and video data. CNNs target to utilise a relatively small number of parameters, which boosts the speed of the model's convergence and makes training easier. There are multiple convolution layers stacked before pooling layers in a typical form of CNNs, analogous to the multi-layer perceptrons (MLPs), followed by fully connected (FC) layers attached to the top layers of the CNNs. The fully connected (FC) layers located on the top of the CNNs are typically represented by artificial neural networks (ANNs) to produce the class label of the input sample. Figure 2.5 depicts a simple structure of a CNN, illustrating its main components.

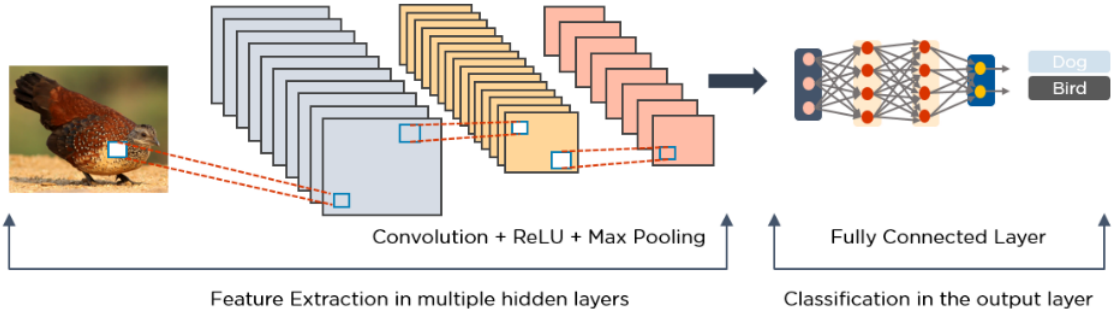


FIGURE 2.5: Simple structure of a CNN model, illustrating its main components. An example of typical CNN architecture with 2 convolutional layers, 2 pooling layers, and a fully connected layer (FC), which provides the final outcome of classification into one of the binary/multiple classes [66].

Artificial neural network (ANN), also called Multi-layer Perceptrons (MLPs), is composed of artificial units of neurons, often referred to as network nodes, designed to mimic the human brain. Three layers of these neurons are arranged next to each other: the input, hidden, and output layers, as shown in Figure 2.6. Each node receives information from data in the form of inputs x, multiplies them using random weights w, and then adds a bias b. Finally, to compute which neuron to be activated, nonlinear functions f (sometimes referred to as activation functions) are used [66]. Linear operations between the input x from the input layer with the parameters w, b are first computed, and an element-wise non-linearity is then applied as follows:

$$f = \eta\left(w^t x + b\right) \tag{2.1}$$

where η refers to a non-linear activation function. Thus, the basis of the CNNs is an MLPs ANN.

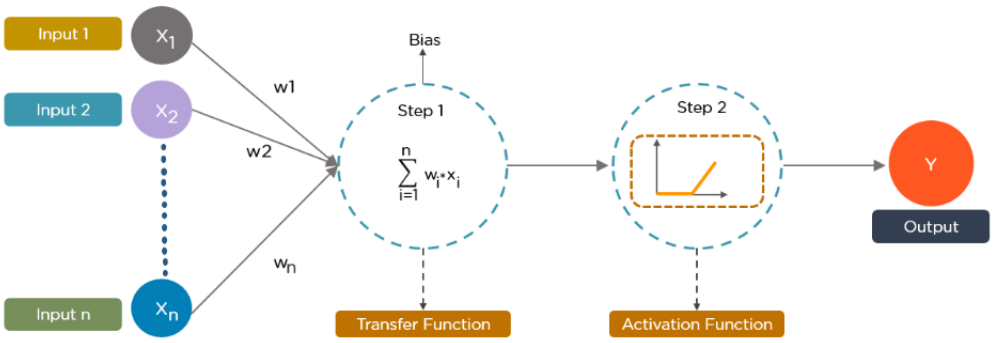


FIGURE 2.6: Artificial neural network (Multi-layer Perceptrons (MLPs)) [66].

Convolutional neural network (CNN) models have demonstrated superior evaluation performance since they can extract features from the data without needing hand-designed attributes. These models include, but not limited to, AlexNet [48], GoogLeNet (Inceptionv1) [67], VGG [68], InceptionResNetV2 [69], Inception3 [70], DenseNet [71], ResNet [72], MobileNet [73], Xception [74], ShuffleNet [75], NASNetMobile and NASNetLarge [76], Darknet-19 [77], EfficientNetB0 [78], Darknet-53 [79], and SqueezeNet [80]. Figure 2.7 shows the relative processing times and accuracy of popular CNN models for image classification, utilising an NVIDIA Tesla P100 GPU and a batch size of 128. Each blue marker in the graph represents the size of model [81].

The utilisation of graphics processing units (GPUs) to train the deep learning models was one of the essential aspects. The first applications for GPUs were in computer gaming. Yet, to construct the back-propagation technique in deep learning algorithms, researchers took advantage of the computing power of GPUs. Due to the acceleration provided by GPUs, researchers were able to train deeper convolutional neural networks, which reduced error rates. Researchers have maintained improving deep learning performance because of theoretical advancements in the field, the use of GPUs, and the accessibility of massive labelled datasets. The usage of GPUs is demonstrated in Figure 2.8 by depicting the percentage of ImageNet entries over time for the image classification task, utilising graphic processor units (GPUs). Through platforms like Kaggle

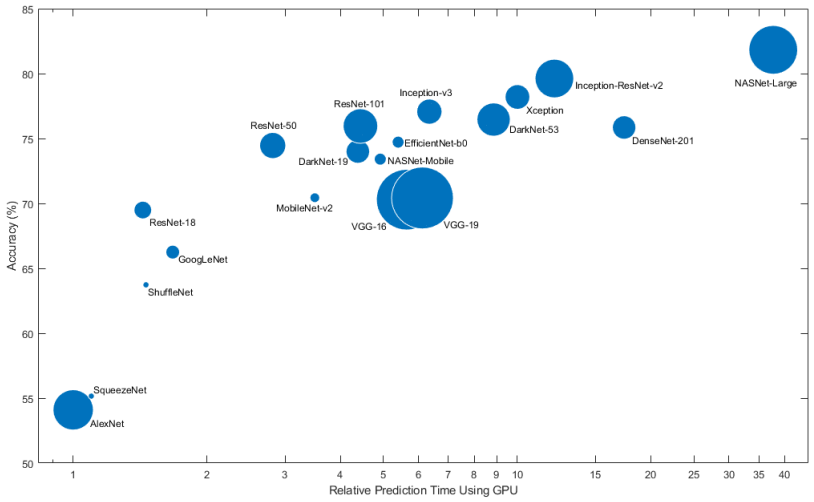


FIGURE 2.7: Most commonly used convolutional neural network models [81].

[82] that provide data and an environment for running deep learning algorithms using GPUs, public data is becoming more conveniently available for analysis [83]. Deep learning, especially CNNs, has become more feasible for challenging computer vision applications due to these breakthroughs.

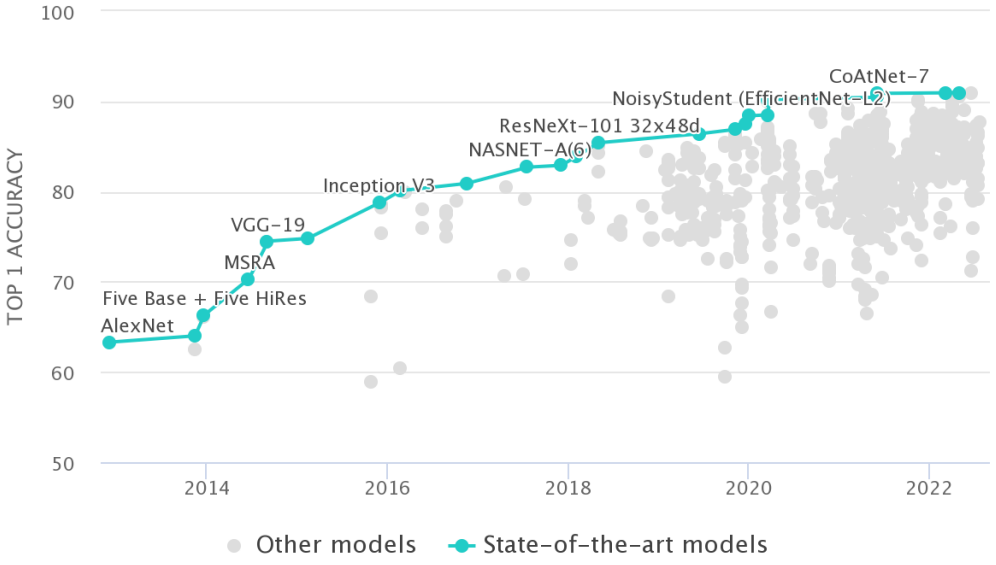


FIGURE 2.8: Percentage of ImageNet entries over time for the image classification task, utilising graphic processor units (GPUs) [84].

In this section, the main architecture of the CNN model is described as follows [85]:

1. **Convolutional Layer**: the fundamental building units of convolutional neural networks (CNNs) are convolutional layers. The convolution operation is the process of applying a pre-defined filter to an input image to produce feature maps. Feature maps highlight the locations and intensity of identified features in an image created by applying the convolutional filter. CNNs are distinctive in their ability to apply several parallel filters on a training dataset, producing significant and representative feature set [86]. The convolution operation is illustrated in Figure 2.9.

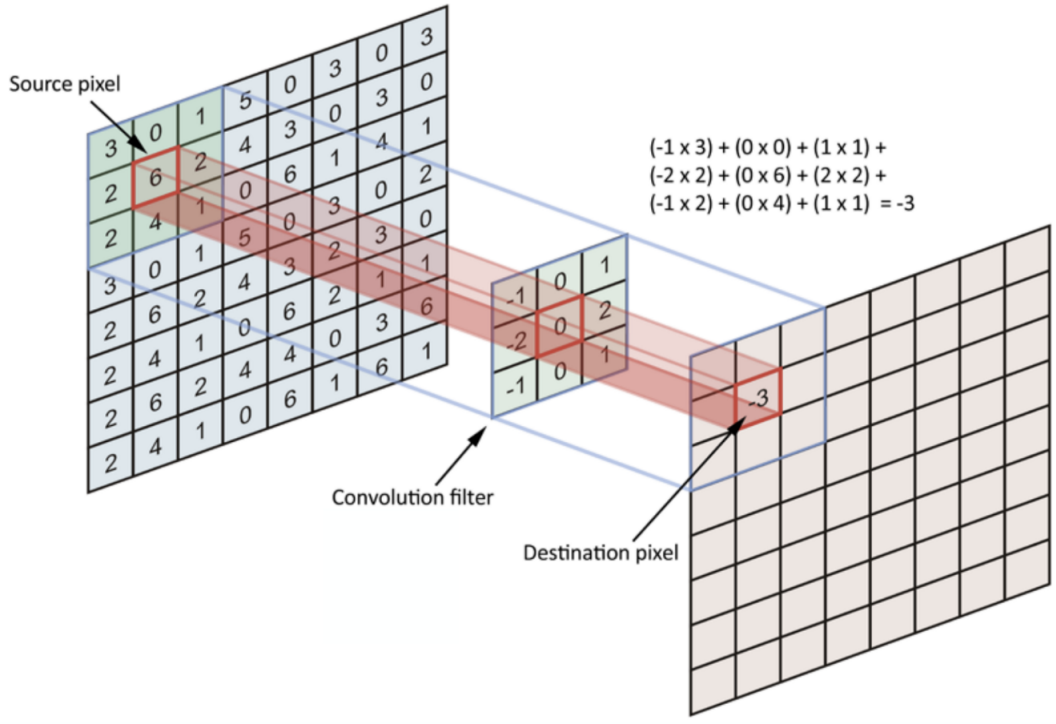


FIGURE 2.9: Explanation of Convolution operation [87].

The convolutional layer needs three hyper-parameters to be predefined. These hyper-parameters are the filter number and size, stride, and zero-padding style. The number of convolutional filters employed in a certain layer is indicated by the filter number. The filter size is the size of the filter's receptive field, typically odd dimensions like 3×3, 5×5, etc. The motion distance of the filter across the image is called stride. The filters slide over the input image 1 pixel at a time if the value of stride is 1. When the stride is two, the filters slide while jumping two

pixels at a time. Padding the input volume with zeros at the edge of the image is a common practice called zero-padding. Zero padding enables adjustment of the spatial output size. The mathematical representation of the convolution operation using k filter weights is given as follows:

$$y_i = f(b + \sum_i x_i * k_i) \tag{2.2}$$

Where f refers to a non-linear activation function, $x_i * k_i$ is a convolution operation between the input x with the convolutional filter (kernel) k_i at position i, and b represents the bias of the model.

2. **Pooling Layer**: downsampling typically referred to as pooling layers, carries out dimensionality reduction and minimises the number of parameters in the input data. The pooling operation slides a filter over the whole input similarly to the convolutional layer, with the exception that this pooling filter lacks weights. Alternatively, the kernel (filter) populates the output array by applying an aggregation function to the values covered by the filter's receptive field. This layer is typically added after the convolution layer. As shown in Figure 2.10, there are two primary forms of pooling operation: maximum pooling (max) and average pooling. The pixel with the highest value is placed in the feature map's output array as the max pooling filter slides over the input image. This method is applied more frequently than average pooling. Sliding over the input, average pooling determines the average value covered within the receptive field and passes it to the resulting output array. Both window size and stride are hyper-parameters for the pooling layer. The pooling layer is a lossy layer that loses information provided in the feature map but offers CNN several advantages. This layer assists in reducing the complexity and dimensions of the feature map and lowering the danger of overfitting [88].

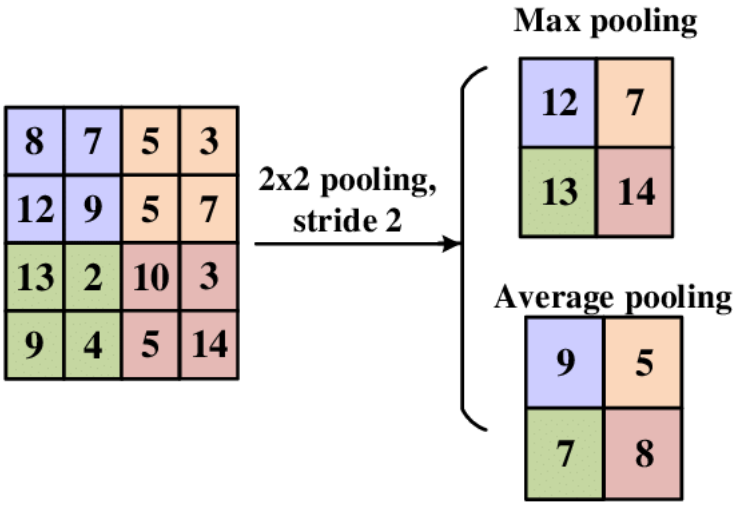


FIGURE 2.10: Explanation of pooling operations: Average and Max pooling layers [89].

3. **Fully-Connected Layer**: in convolutional neural networks, the convolutional and the pooling layers together make up a block. Depending on the task's complexity, the number of these layers could be expanded to capture finer attributes (features) at the expense of additional computation power. The resulting feature representation is flattened and given to a typical fully-connected neural network for image classification/regression. FC layers are typically used to optimise objectives like class scores and are often attached to the top of the CNN layers [90].

4. **Activation Functions**: a neural network's activation function describes converting the weighted sum of the input image into one neuron or multiple neuron outputs. A "transfer function" is another name for the activation function. It may be referred to as a "squashing function" if the output range of the function is constrained. Numerous activation functions have nonlinear behaviour, known as "nonlinearity". Various activation functions may be employed in different locations (layers) of the network architecture. The selection of activation functions significantly influences the network's performance. It is also common for activation functions to be differentiable, which enables the computation of the first-order derivative for a provided input value. Given that the CNN is typically

trained using the back-propagation, which needs the derivative of the estimated error to update the model's weights, the activation function should be differentiable. Although CNN uses many different kinds of activation functions, only a few of them are performed for real-world scenarios. Examples of activation functions utilised in the CNN are Tanh, Sigmoid, ReLU, Leaky ReLU, and Noisy ReLU [91]. ReLU is the most commonly utilised function in the CNN context, which is defined as follows:

$$\theta : x \rightarrow max(0, x) \tag{2.3}$$

5. **Loss Functions**: the disparity between the output generated by the CNN and the true (target) value is measured by the loss function. The gradients needed to update the weights can be derived from the loss function. The cost is determined by taking the average of all losses over training data fed to the network during the training phase. Several loss functions, such as Binary Crossentropy (BCE), Categorical Crossentropy(CCE), and Mean Square Error (MSE), can be utilised to achieve the objective of the learning process. The probability for each class of the data used for training is produced by Softmax, where the sum of these probabilities should be equal to one. A Softmax activation function with a Cross-Entropy loss makes up Softmax loss. The binary classification applications typically employ BCE loss. When utilising the BCE loss function, only one neuron unit in the output layer is needed to classify the dataset into two categories. In CCE, for multi-class classification problems, the number of (nodes) neuron units on the output layer must be equal to the number of classes provided in the training dataset. Additionally, a softmax activation should be used within the final layer such that the output of each neuron maintains a probability score between (0–1). Regression tasks usually employ MSE loss. As the name implies, this loss is determined by averaging the squared discrepancies between ground truth (real value) and predicted outcomes. For instance, BCE can be defined as follows:

$$H_{\hat{d}}(d) = -d\log(\hat{d}) - (1-d)\log(1-\hat{d}) \tag{2.4}$$

Where $\hat{d}$ represents the estimated outcome from the network and d refers to the target value.

6. **Over-Fitting and Regularisation**: model over-fitting is a serious problem and can cause the model to produce misleading information. When the CNN algorithm works exceptionally well on training data but keeps failing on unseen test data, this condition is referred to as over-fitting. In contrast, the under-fitted models do not learn enough from the provided training dataset and subsequently do not train well. The models are deemed to be "appropriate-fitted" if they work well on both the training and unseen testing sets of data. Figure 2.11 shows the phenomena of over-fitting, under-fitting, and the correct fit.

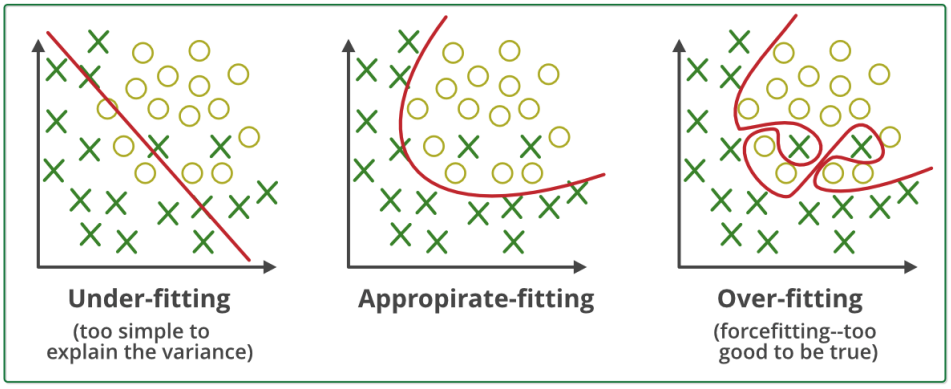


FIGURE 2.11: Over-fitting, under-fitting, and the correct fit in a model.

The most common technique typically used to overcome over-fitting is regularisation. Regularisation, in general, penalises the coefficients that cause the over-fitting of the model. Deep learning researchers use many concepts and techniques to help reduce the model regularisation, including weight decay ($L1$ regularisation (LASSO) and $L2$ regularisation (Ridge)), dropout, data augmentation, transfer learning, early stopping, and batch normalisation. These techniques are described as follows [85]:

- Weight decay: the names $L1$ and $L2$ regularisation come from a vector's corresponding $L1$ and $L2$ norms. When computing the cost using a loss function, an auxiliary term known as regularisation is introduced to penalise complicated CNN models. Both $L1$ and $L2$ might apply a penalty to the cost function considering the CNNs' complexity.
- Early stopping: automatically halt the training process when a particular performance metric (such as accuracy/validation loss) stops progressing.
- Data augmentation: it is a method for producing new data from the existing set to artificially increase the size of the data used for training. Using data augmentation to reduce the over-fitting is a reasonable solution if the size of the training dataset is not enough to achieve improved performance from the CNN model. Geometric transformation, which allows to flip, rotate, crop, or translate images arbitrarily, is one example of data augmentation strategies. Changing the colour channels of images or enhancing colour are examples of colour space transformation which target increasing the image data based on the data augmentation concept. Furthermore, applying sharpening and blurring filters on images is also a data augmentation technique.
- Transfer learning: in the transfer learning scheme, a previously trained model is used as the foundation for a new model on a similar computer vision task. Simply said, a model developed for a certain task is used for another similar task to facilitate optimising and training the second task easily and quickly. Transfer learning enables to utilise of data from bigger datasets to reduce over-fitting brought on by small datasets.
- Batch normalisation: every layer of the network may learn more independently thanks to the layer of batch normalisation. The output of the earlier layers is normalised using batch normalisation. Normalisation is a pre-processing method applied to normalise data. Normalisation helps to process data from different resources and adapt it to belong to the same range.

CNN may experience issues, making it significantly more challenging to train and slowing down its learning if the data is not normalised. Instead of applying normalisation on the raw data, a CNN does it between the layers using a batch normalisation layer. Rather than using normalising the entire dataset, the normalisation is performed in mini-batches.

- Dropout: is a deep learning approach where nodes (neurons) in CNN are removed or dropped out to mimic training many architectures at once. Notably, dropping out can significantly lower the over-fitting occurrence. When dropout is performed, a "thinned" model is produced with distinct combinations of the network nodes being deleted at random intervals throughout the training process. According to a probability hyper-parameter p, a new thinning CNN model is generated every time the model's gradient is adjusted. It is possible to think of training a network using the dropout technique as training many individual thinned networks and combining them into one model that inherits the salient features of each thinned model.

7. **Weight Initialisation, Optimiser Selection and Model Learning**: typically, to develop a deep learning algorithm, the network architecture should be initially established, and then training is performed to learn the parameters. The typical learning procedure in the CNN includes: (1) weight parameter initialisation, (ii) establishment of an optimisation strategy, and (iii) carrying out the following steps iteratively: (a) propagating an input in forwarding pass, (b) the cost function is computed, (c) the gradients of the cost are computed using back-propagation, and (d) the parameters are updated using the gradients based on the optimisation technique. Initialising the network parameters is the fundamental step that must be taken into account while creating the network. If the initialisation is performed correctly, optimisation will be accomplished in the shortest time; otherwise, utilising gradient descent to converge to the minima would not be feasible. One of the most commonly employed weight initialisation techniques

can be given in Eq. (2.5) [88]:

$$w \sim \alpha.\upsilon[-\delta, \delta] + \beta.\eta(0, \delta) + \gamma \quad with \quad \alpha, \beta, \gamma \geq 0 \tag{2.5}$$

"Where the term $\eta(0, \delta)$ represents the normal distribution with mean zero and variance of δ and the term $\upsilon[-\delta, \delta]$ represents the uniform distribution" [88]. The values of $delta$, $alpha$, and $beta$ parameters have all been defined using a variety of techniques such as Xavier/Glorot (uniform) [92], Xavier/Glorot (normal) [92], and He [93]. The learning process in deep learning algorithms is iterative. Thus, it is crucial to rapidly train the model to finish the iterative cycle as soon as possible because there are many parameters that need to be adjusted and optimised. Stochastic Gradient Descent (SGD) is a popular optimisation approach in CNN models. A number of optimisation methods including, Momentum [94], AdaDelta [95], AdaGrad [96], Adam [97], and RMSProp [98], were proposed based on SGD. All these techniques enable learning of the network, but in terms of speed, some techniques outperform others.

2.5 Computer-Aided Diagnosis Systems: An Overview of Automated Skin Lesion Image Analysis Methods

Computer-aided diagnosis (CAD) systems are frequently used to identify and categorise skin diseases. Such systems, which have high detection results, greatly minimise the time and effort spent by clinicians. There are some difficulties involved in putting up such systems. Two sorts of techniques are suggested in the literature, and they are directly correlated with the dataset size of skin lesion images. As a result, machine learning-based algorithms (ML-based) are frequently utilised for small datasets, beginning with the identification of the lesion region because the areas under analysis

may include noise. An accurate lesion diagnosis will enable the extraction of pertinent features and attributes, ultimately leading to a high recognition rate. On the other hand, where numerous architectures have been presented, deep learning-based techniques (DL-based) are the most popular and effective for handling large datasets. The strengths of machine learning-based algorithms (ML-based) and deep learning-based algorithms (DL-based) models have recently been combined in several interesting hybrid concepts. This section presents the skin lesion analysis methods developed for different computer vision tasks, including classification, segmentation, and detection.

2.5.1 Machine Learning Based Approaches

The most promising basis for an automated computer-aided lesion diagnosis is digital dermoscopic images [99]. The two basic goals of automatic dermoscopy image analysis are to recognise dermoscopic characteristics in each image and link those feature findings to the diagnosis. A typical automated dermoscopy image analysis pipeline includes preprocessing, lesion segmentation, feature extraction, and, finally, classification. The following is a short description of these steps:

1. **Preprocessing**: dermoscopy images undergo various preprocessing steps, such as image enhancement (i.e. shading removal, colour correction, and contrast adjustment), the transformation of colour space, and artefact removal. The artefact removal may include removing ruler markings, hairs, black frames, air bubbles, and ink markings [100–102].

2. **Lesion segmentation**: includes separating skin lesions from the surrounding healthy skin. There are two reasons why lesion segmentation is crucial. The lesion border, in the first place, offers critical information for a precise diagnosis. For instance, irregularity or asymmetry at the border might indicate cancer. Second, it is standard procedure to exclude the healthy skin around the lesion by doing feature extraction solely on the lesion. Therefore, the precision of the segmentation determines how representative an image's attributes are. For various

reasons, including the existence of artefacts and the lack of contrast between the lesion area and its surroundings, lesion segmentation imposes many challenges making it one of the most investigated topics. To obtain insight into the lesion segmentation methodologies, the segmentation techniques can be categorised into two kinds, low-level and high-level strategies. Low-level strategies are standard and classical methods that necessitate post-processing. They are also faster and simpler computationally, such as threshold based methods [103, 104], region based approaches [105], and edge based approaches [106]. High-level segmentation methods combine low-level methods to develop more comprehensive and sophisticated segmentation techniques, eliminating the need for postprocessing and handling poor-contrast borders such as soft computing based approaches [107], fusion based techniques[108], and deformable models [109].

3. **Feature extraction**: entails the recognition, characterisation, and interpretation of visual features in an image. Although the term ”image feature” has a fairly broad definition, it may be perfectly described as generating abstractions of information from an image that is significant for diagnosis. Depending on how the algorithms function, these features are generally handled at various levels, which are low-level features and high-level features. For instance, the distribution of image intensity, colour, or texture might be employed as a low-level feature to solve a computational problem directly tied to diagnosing a health condition. The features extracted from the dermoscopic images include ABCDE rule-based features [110], seven Point checklist based features [111], CASH algorithm based features [112], shape features [113], colour features [114], texture features [115], and high-level features [116].

4. **Classification**: the final phase is often classification. Lesion classification produces various results depending on the application, but frequently it yields an estimation of the likelihood of malignancy. A common strategy for achieving this goal is typically harnessing the traditional scheme of a supervised learning

fashion, which relies on training data. Choosing a subset of pertinent characteristics (features) is sometimes used in the model building before the classification phase, which is typically called feature selection. There are several well-known classifiers adopted in melanoma diagnosis CAD systems including, KNN based methods [117], support vector machines (SVM) [118], decision trees [119], ensemble classifiers [120], logistic regression [121], Bayesian network [122], and artificial neural networks (ANN) [123].

2.5.2 Deep Learning Based Approaches

It is evident that deep learning, particularly convolutional neural networks (CNNs), has taken the lead in solving many computer vision applications [124] since Krizhevsky et al. [48] achieved the top rank in ImageNet competition held in 2012. With an ever-growing number of applications that employ deep learning technology (DL) to assess medical conditions using images captured from different modalities, the field of medical imaging has also adopted this concept [43]. To the best of my knowledge, Codella et al.'s publication in 2015 [54] was the first article that adopted CNNs for dermoscopic image data analysis. In recent years, researchers have widely investigated DL models to discover appropriate features and obtain precise diagnostic performance [125]. The ISIC challenge held in 2017 [126], in which twenty-two out of twenty-three submissions employed the CNN models, provides evidence of the prominence of DL in the dermoscopy discipline.

Dermoscopic image datasets have been widely used in several deep learning-based research studies for lesion diagnosis and identification. In [127], a deep learning algorithm, InceptionV4 CNN, was trained on a large dataset made up of (100,000) dermoscopic image samples of two classes: melanoma and benign tumours. Haenssle et al. [127] compared their results with 58 dermatologists. Dermatologists achieved an average specificity score of 71.3 % and a sensitivity score of 86.6 % on (100) dermoscopic images (25% melanoma and 75% benign). In contrast, the CNN approach

attained a specificity score of 63.8 % and a sensitivity score of 95 %. Tschandl and his colleagues [128] assessed the average performance of 511 human readers and AI systems on 1511 test images. A total of 139 AI algorithms were developed by machine learning labs that took part in the ISIC 2018 competition. In terms of outcomes, AI systems outperformed human readers in terms of correct assessment and lesion diagnosis. To diagnose non-pigmented skin malignancies, Tschandl and his colleagues [129] applied the well-known deep learning models ResNet50 and InceptionV3 CNNs to 7895 dermoscopy images. Ninety-five dermatologists are grouped into three teams of experts depending on their experience to compare performance. With novice and intermediate group levels, the CNN models outperformed human groups in terms of accuracy and were on a level with human specialists.

A deep learning model, ResNet50 CNN, was used by Brinker et al. [130] and its performance was compared to 157 dermatologists at twelve University hospitals in Germany using 100 dermoscopy images, which included 80 nevi and 20 instances of melanoma. On the dermoscopic dataset, dermatologists obtained a sensitivity score of 74.1% and a specificity score of 60.0%, while the CNN model obtained a sensitivity score of 84.2% and a specificity score of 69.2%. Maron et al. [131] tested the specificity and sensitivity of the ResNet50 CNN model against 112 dermatologists for the multi-class identification of skin diseases, including nevi, melanoma, SCC, BCC, and benign keratoses. The deep learning technique significantly outperformed dermatologists. Dermatologists and the InceptionV4 CNN model were compared by Haenssle et al. [132] using 100 instances (40 malignant and 60 benign samples). The dermatologists' average specificity and sensitivity scores were 80.7% and 89%, respectively, compared to the deep learning algorithm's 76.7% and 95%. According to [133], other deep learning methods, including the MobileNet and Long Short-Term Memory (LSTM), have also been proven to be efficient for skin lesion identification. Authors of [134] introduced DUNEScan (Deep Uncertainty Estimation for Skin Cancer). This web server examines the uncertainty in widely used convolutional neural network-based skin cancer identification models (CNNs). Recently, the authors of [135] presented a comprehensive

survey of the latest DL algorithms applied for analysing skin lesions and diagnosing skin cancer.

To segment the lesions and extract areas of interest, existing segmentation models are modified to achieve desirable performance. The authors of [136] suggested a tweaked U-Net variant. The performance of U-Net in lesion segmentation was enhanced by fusing DenseNet and ResNet. Context modules with dense connections are intercalated between the encoder's convolutional layers. Similar to this, localised modules are intercalated between the upsampling layers of the decoder. 3D CNN is a new trend in segmentation architectures that is able to produce more reliable lesion segmentation findings. For instance, to obtain a more precise lesion segmentation for the identification of melanoma from hyper-spectral pathology images, the authors of [137] presented a 3D CNN called HyperNet. The conventional convolution and the dilated convolutional filters for multiple-scale features were merged in their proposed architecture. A fusion path is established between the blocks of the encoder and decoder. A residual learning scheme was developed to improve training effectiveness and was motivated by V-Net [138].

Additionally, there has been a range of deep learning-based research on lesion identification utilising clinical images that are frequently taken with mobile cameras of various skin lesions for examination and integration into patients' medical records. Yang et al. [139] accomplished skin lesion identification clinically based on the ABCDE rule. They evaluated how well the suggested strategies performed compared to deep learning techniques and dermatologists. In terms of accuracy, it scored 57.62%, beating out ResNet, the top-performing deep learning technique, which scored 53.35%. Only senior physicians with extensive expertise in skin conditions outperformed in terms of accuracy, topping at 83.29% on average. On several publicly available datasets and twelve skin conditions, Han and his colleagues [140] learned the DL model (ResNet-152 CNN) to categorise the images into multi-class labels (twelve conditions). Using 480 randomly selected images, the algorithm performed equally with the team of 16 dermatologists while outperforming them in diagnosing BCC. On 100 clinical images

made up of 20 melanoma and 80 nevi samples, Brinker et al. [141] assessed the performance of the ResNet50 CNN against 145 doctors. The deep learning approach obtained the same sensitivity and a higher specificity value of 69.2% compared to dermatologists, who had an average specificity score of 64.4% and a sensitivity score of 89.4%. Using 6009 clinical image data of 14 skin conditions, covering both malignant and benign pathologies, Fujisawa et al. [30] explored a DL approach. Using 140 test images, the deep learning model outperformed 9 dermatology trainees (reported accuracy of 41.7%) and 13 senior dermatologists (reported accuracy of 59.7%), achieving a diagnostic accuracy of 76.5%.

2.6 Summary

In this chapter, an overview of skin lesions was reported by highlighting the skin lesion types and causes (covered in Section 2.1). Furthermore, skin lesion modalities presenting the imaging techniques in dermatology were described in Section 2.2. The diagnosis techniques of skin lesions clinically were reported in Section 2.3. Moreover, Section 2.4 provided a review of artificial intelligence, learning methods, deep learning, computer vision, CNN layers and models. Finally, existing computer-aided diagnosis systems, including an overview of automated skin lesion analysis methods, were described in Section 2.5. It is abundantly apparent from the earlier research studies discussed in this chapter that deep learning and computer vision have been shown to perform well when it comes to analysing skin lesions. The inspiration to employ Convolutional Neural Networks (CNNs) in the analysis of skin images to design computer-assisted diagnostic systems, leveraging various deep learning techniques and skin imaging modalities, has been fostered in this thesis. The next Chapter, Chapter 3, presents a regression-based method for grading skin lesions related to acne from facial images. The following two chapters, Chapter 4 and Chapter 5, present automated methods developed for detecting and classifying the skin lesions related to melanoma.

Chapter 3

Grading of Skin Lesions Related to Acne From Facial Images using Regression-based Deep Learning Model

The work presented in this chapter describes a regression-based deep learning model developed to assess and grade the severity of acne lesions using facial images. The developed model reported in this chapter was published in Alzahrani et al. [142]. The contribution of the first author to work presented in this chapter is the conceptualisation of the idea, proposing the methodology, development of the model, data and results analysis and writing up. The labelled dataset used for training and testing was provided by the author of [143].

3.1 Introduction

Acne vulgaris is simple to diagnose; however, its polymorphic structure makes it difficult to assess its severity. A physician's validated assessment generally determines the effectiveness of acne treatment. For assessment by the physician, the different acne

lesion types involve being counted independently. Acne affects about 80% of adolescents [144], with 3% of men and 12% of women experiencing symptoms even through adulthood [145]. As a result, there are a large number of acne patients who require immediate treatment, as acne can cause scars and pigmentation as well as a sense of inferiority and depression [146]. Dermatologists need to know the severity of acne to make a precise and appropriate treatment selection [39]. However, due to the limited time available for consultation, the manual validated evaluation of acne might be difficult and time-consuming. Additionally, junior dermatologists need a reference diagnosis that is objective and trustworthy. With the development of imaging modalities, widespread availability of digital cameras, and deep learning (DL) techniques, automatic acne detection and severity evaluation systems from photographs would help dermatologists attain a more reliable and consistent assessment of acne in clinical practice trials [43, 47, 147].

3.2 Related Work

Remarkable progress has been made for automated acne lesion analysis in recent years covering several acne lesion analysis tasks such as acne classification [148–151], segmentation [152–155], detection and localisation [150, 153, 154, 156, 157], and severity grading [143, 154, 158–161]. The analysis of acne lesions was accomplished by image processing techniques [153, 155], extracting hand-crafted features and passing them into a classier model [150, 154], and automated feature learning using CNNs [149, 157, 160]. In this work, we address the problem of acne severity grading from facial images.

Several methods have been proposed in the literature targeting the automated severity grading of acne lesions. In [154], hand-engineered features were extracted from segmented acne areas and passed into an SVM model to classify the severity of acne lesions into four levels following the criteria established by Ramli [162]. Their method was assessed on a private dataset composed of 35 images. Alternatively, the authors

in [158–161] exploited CNNs to extract the features automatically and subsequently, graded the severity of acne lesions following the criteria established by IGA (three levels) [163], Hayashi (four levels) [39], GEA (five levels) [164], and IGA (five levels) [163], respectively. Those developed systems were trained and evaluated on private datasets consisting of 472, 4700, 5972, and 479 images, respectively. The authors in [143] presented acne counting and grading method based on the label distribution learning paradigm (LDL) with CNN to classify the acne severity into four levels following Hayashi assessment criteria [39]. They evaluated the performance of the developed method on a public dataset of 1457 images. However, the performance of these developed approaches has limitations and experiences challenges. The performance of handcrafted feature regression-based methods highly depends on the type of features extracted from a specific dataset. Furthermore, those features might be applicable in a particular dataset but may not generalise well on other datasets. On the other hand, CNN regression-based methods globally estimate outcomes from features without concerning the detailed location of understudied acnes that should be considered following the grading criteria.

To tackle the aforementioned limitations, we developed a new computer-assisted image analysis approach to grade the severity of acne lesions called dilated UNet dense regressor guided by an attention mechanism. Inspired by the scenario of crowed counting from kernel density maps [165, 166], region of interest density maps for acne lesions are generated to produce the count of lesions within a particular area of interest. Thus, we propose a method to count objects of interest, represented by acne lesions, and, subsequently, grade the severity of acne in facial images. Following [167], we adopt a fully convolutional UNet, which is originally used for segmentation, to construct the regressor responsible for generating the density maps. In addition, following [168], we exploit the multi-scale dilated filters to implement the bottleneck convolutional filters of UNet. Accordingly, we developed a multi-scale dilated UNet regressor for density map generation. The proposed convolutional network module uses dilated convolution filters to systematically aggregate multi-scale contextual information trying to mitigate

the loss in resolution. On the top of the multi-scale dilated UNet regressor, we embed the prior information of bounding boxes as an attention mechanism generated by Faster R-CNN [169], which was originally developed for object detection. In this fashion, we merge the dilated UNet dense regressor with the Faster R-CNN network for density map regression, allowing us to determine the count of acne lesions and subsequently grade the severity.

Beyond the bounds of acne lesion counting, the concept of object counting has been widely applied in a variety of scenarios, including cell counting in microscopic images [170], tree counting [171], animal counting [172], vehicle counting [173], and crowd counting [174]. Generally, estimating the number of any objects in a still image or a video is typically defined as a counting problem. The object counting methods can be broadly divided into two categories: detection and regression-based techniques. The counting-by-detection approaches, which use detectors to detect each object in an image or video, were widely used in early efforts addressing the object counting topic. To extract low-level features, these approaches require well-trained classifiers such as HOG, histogram-oriented gradients [175], and Haar wavelets [176]. Recent approaches leveraging CNN-based object detectors to achieve end-to-end learning paradigms, such as YOLO3 [79], SSD [177], and Faster R-CNN [169], have considerably improved counting accuracy.

Different from counting by detection, regression-based approaches obtain the count without explicitly detecting and localising each object. Global regression and density estimation are the two types of regression-based counting techniques. Global regression methods [143, 156] explicitly predict the final count from images by learning the mapping between image features. In contrast, density estimation-based methods [165, 178] first estimate a density map, which is then integrated (summed) to produce the final count. Density estimation typically outperforms global regression because it makes use of more spatial information about objects in an image. However, acne lesion counting based on either regression or detection approaches is insufficient to handle both high-

and low-density regions of acne lesions simultaneously. When counting using regression solely, there is a risk of overestimation when there are low densities of objects (sparse regions). Similarly, counting by purely detection methods would result in the underestimation problem on occasions with high densities of objects (dense regions). Thus, counting by detection performs comparably better in the sparse regions; on the other hand, counting by regression performs comparably better in the dense areas [179]. This motivated us to establish a system that takes advantage of regression (Dilated UNet Regressor) potentials and impressing attention to the acne lesion positions detected by the detecter (Faster R-CNN), inspired by [180, 181].

3.3 Materials and Methods

3.3.1 Materials

To conduct the experiments in this research work, a publicly available dataset named ACNE04 is used [143]. The number of lesions and global acne severity are annotated in the ACNE04 dataset by specialists. Images of acne lesions are collected using a digital camera with patients' consent when physicians are making a diagnosis. Images are taken at a 70-degree angle from the front of patients to meet the requirements of the Hayashi grading criteria [39]. The specialists then manually annotate the images using the annotation tool provided. The ACNE04 contains 1457 images of lesions with 18,983 bounding boxes.

3.3.2 Methods

In this section, we describe the proposed attention-guided UNet dense regressor for addressing the task of acne counting and severity grading in detail. The developed architecture incorporates dilated UNet dense regressor for density regression with the information of bounding boxes generated from the Faster R-CNN network, producing a hybrid detection–regression framework. Figure 3.1 presents the abstract level of the

proposed architecture for acne severity grading. In the following sections, we will describe the phases of the proposed model, including the ground truth generation of kernel density maps, bounding boxes generation by Faster R-CNN, and the construction details of the dilated UNet dense regressor guided by the attention mechanism.

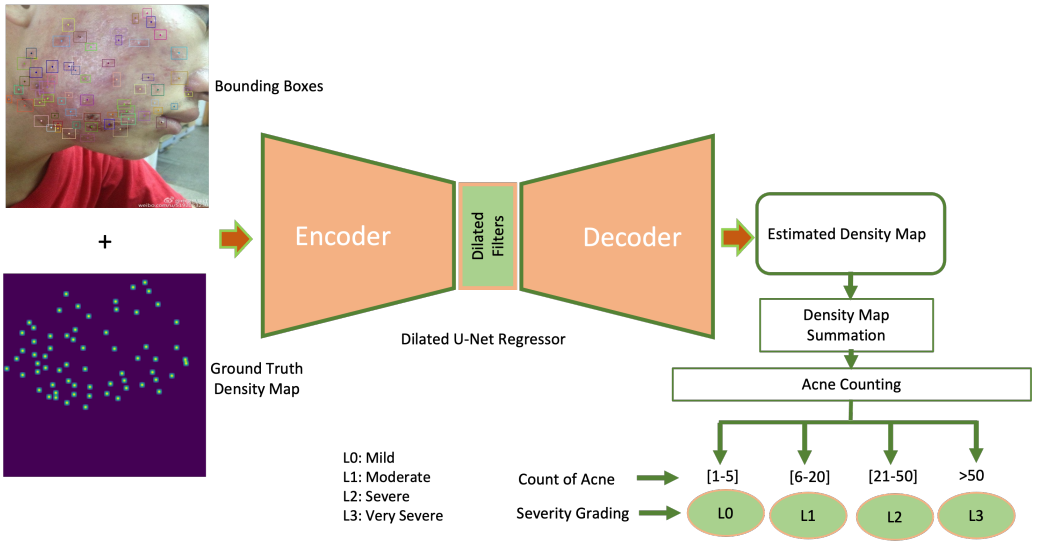


FIGURE 3.1: Block diagram of proposed acne counting and grading system.

Generation of Ground Truth Kernel Density Maps

Due to severe overlapping and variation in the size of the acne, individual acne detectors might encounter problems in locating facial skin lesions in dense regions. Hence, the challenge of acne counting is handled as estimating a kernel density function whose integral over each image region yields the number of acne in that image. Thus, the resulting density map would preserve information indicating the presence of lesions in a specific area. To estimate the acne density map from an input facial image, the UNet density map regressor is first trained on training facial images along with their ground truth density maps. The quality of generated ground truth density map for a given training image determines the performance of the developed method. To generate a map of acne density for training data, it is required to provide point annotations for acne lesions. As the data used in this study were provided with bounding boxes around each acne, we first determined the centre point value of each bounding box around acne lesions producing dot-annotation pixels in an image. The centre point coordinates

(c_x, c_y) are formulated by the value of the top left corner of the boxes (x_{min}, y_{min}) and bottom right corner (x_{max}, y_{max}). The centre point coordinate of a bounding box in an image is introduced using the following formula:

$$[c_x, c_y] = (x_{min} + \frac{x_{max} - x_{min}}{2}, y_{min} + \frac{y_{max} - y_{min}}{2}) \tag{3.1}$$

To generate the density map $F(x)$ given a point value of pixel x_i located at $[c_x, c_y]$ from total R point-annotation acne lesions in an image, the method for generating density maps used in [165] was followed. This can be achieved by convolving $\delta(x - x_i)$ with Gaussian kernel G_σ. The Gaussian kernel is set with a fixed spread parameter σ of 4 and kernel size of 15 by blurring each acne annotation point as follows:

$$F(x) = \sum_{i=1}^{R} \delta(x - x_i) * G_\sigma(x) \tag{3.2}$$

The choice of a Gaussian kernel with a standard deviation of 4 and kernel size of 15 for generating density maps in counting is often used due to its desirable properties [165, 174, 178, 182]. The Gaussian kernel is used because it is a smooth, bell-shaped function that models the spread of objects in a crowded region. The standard deviation of 4 determines the spread of the Gaussian, and a larger standard deviation results in a wider and smoother density map, while a smaller standard deviation results in a narrower and more focused density map. A standard deviation of 4 is often used because it provides a good balance between accuracy and robustness, allowing the density map to capture the general distribution of objects in a region while ignoring small variations and noises in the image. The kernel size of 15 determines the size of the Gaussian kernel in pixels, and a larger kernel size results in a larger and more blurred density map, while a smaller kernel size results in a smaller and sharper density map. A kernel size of 15 is often used because it provides a good balance between spatial resolution and accuracy, allowing the density map to capture the general distribution of objects in a region while preserving enough detail to accurately estimate the number of individuals in the image. In summary, the choice of a Gaussian kernel with a standard deviation of

4 and kernel size of 15 for generating density maps is often used because it provides a good balance between accuracy and robustness, and between spatial resolution and detail preservation, making it a suitable choice for object counting.

Bounding Boxes Generation by Faster R-CNN

Faster R-CNN [169] adopted in this work for generating bounding boxes of acne lesions has two networks: a region proposal network (RPN) for generating region proposals and a network that uses the generated proposals to detect objects. There are two heads on the top of the Faster R-CNN, one for object classification and another for bounding boxes regression. ResNet50 model [72] is used as a backbone of the Faster R-CNN detector for feature extraction. The block diagram of Faster R-CNN is illustrated in Figure 3.2. The principle work mechanism of the adopted detector is as follows: i) the RPN network generates region proposals, ii) for all region proposals resulting in the image, a fixed-length feature vector, features extracted by ResNet50, is obtained from each region using the ROI Pooling layer and then classified as object or non-object, and finally, iii) the class scores of the predicted objects in addition to their bounding boxes (BB) are given.

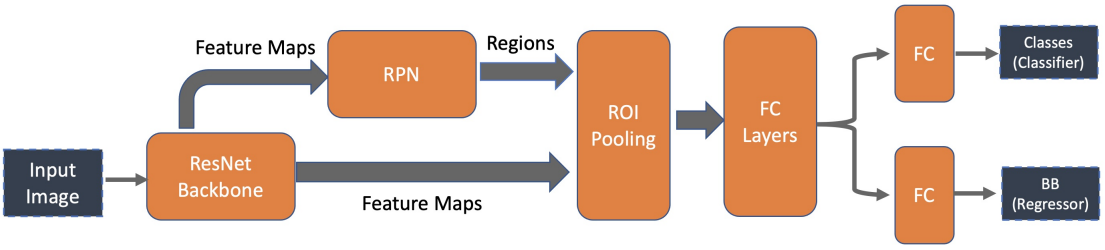


FIGURE 3.2: Block diagram of Faster R-CNN.

Dilated UNet Dense Regressor Guided by Attention Mechanism

The overall structure of the proposed dilated UNet dense regressor architecture with attention module is shown in Figure 3.3. The UNet encoder–decoder segmentation model has been adapted by integrating bounding box information at the level of the skip connections. The outcome of the bounding boxes acts as an attention assistant

module. Using element-wise multiplication of the elements located inside the region of interest of the bounding boxes, feature maps extracted at different scales from the contraction path are fused with features extracted from bounding boxes and then passed to the expanding path.

When it comes to adding attention blocks at the skip connection level, the Attention-UNet developed in [183] is similar to the proposed model by inserting convolutional filters in the middle of the encoder and decoder paths. However, the structure of attention models used to focus on relevant features as well as the strategies to which each model establishes the constraints differ considerably. While bounding boxes were utilised in the proposed dilated UNet dense regressor architecture to guide the network on where to seek through the network until reaching the bottleneck, Attention-UNet [183] employs inputs provided by the bottleneck output and moves upward through the skip-connections. Inserting the convolutional filters in the middle of the encoder and decoder paths in the dilated UNet dense regressor model helps the model adjust what it learns by concentrating on the attention areas. This results in the enhancement of feature detection within specific regions of the facial image.

The details of the proposed dilated UNet dense regressor architecture is shown in Table 3.1. UNet [167] is a segmentation network architecture built upon fully convolutional neural networks (FCNs). Unlike FCNs, UNet adopts the symmetry structure of encoder and decoder (contraction and expanding paths). The UNet architecture consists of three sections: the contraction, the bottleneck, and the expansion section. UNet's contracting path (shown on the left in Figure 3.3) is similar to that of a standard CNN, with a combination of convolutional and max-pooling layers. It gradually decreases the size of feature maps while increasing the number of feature channels, allowing the model to learn both global and local features. The output size of the encoder path (contracting path) passing to the bottleneck is 1/16 of the original input size. If convolutional and pooling layers have been kept adding to the bottleneck, the output size would be further downsized, making it difficult to produce high-quality density maps. Inspired by the work [168], dilated convolutional layers are deployed in the bottleneck to extract

more salient information while preserving the output resolution. A small-size kernel with a $k \times k$ filter is typically enlarged to $k + (k-1)(r-1)$ with a dilated stride parameter r in dilated convolution scheme. As a result, it enables flexible aggregation of multi-scale contextual information while maintaining the same resolution. A 2-D dilated convolution can be formulated as follows:

$$y(m,n) = \sum_{i=1}^{M} \sum_{j=1}^{N} x(m + r \times i, n + r \times j) w(i,j) \tag{3.3}$$

where $y(m,n)$ is the resulted dilated convolution from input $x(m,n)$ and filter weights $w(i,j)$ with the dimensions M and N, respectively. The parameter r represents the dilation rate. If the dilation rate $r = 1$, a dilated convolution returns back into a standard convolution. The third section of UNet, the expanding path (the right part in Figure 3.3), contains a succession of convolution and deconvolution components that can step-wise up-sample the feature maps to their original size and minimise the feature channels. The skip connections between the contracting and expanding paths combine and concatenate features from both sides, forcing the model to collect both local and global information.

This dilated UNet dense regressor is augmented with features of the parallel bounding boxes generated by Faster R-CNN in the skip connections between the encoder and decoder segmentation model. This helps to embed bounding box information as an attention mechanism for acne lesions at different scales in the model. The regression-based model (UNet) works well on dense acne lesions on the facial images, whereas the detection-based model (Faster R-CNN) provides better detection of sparse acne lesions. Thus, integrating the detection attention model in one framework with a regression model helps guide and bring the attention of the regressor to the sparse acne lesions that the dense regressor could miss. The bounding boxes are fed to two convolutional layers (attention module) for location feature extraction. The bounding boxes provided to the attention model are a binary map representing the attention region that corresponds to the location of the acne lesions. The intersection of the un-pooled map from a level contracting layer and the feature map of acne lesions from the attention module is

produced and concatenated with the features from the up-sampling layers within each skip connection. Finally, a 1×1 convolutional layer is applied to map the resultant feature vector to the density maps. The difference between the predicted density map and the ground truth is estimated using Euclidean distance. The following is the definition of the loss function:

$$L(\Theta) = \frac{1}{2B} \sum_{i=1}^{B} \| Z(X_i; \Theta) - Z_i^{GT} \|_2^2 \tag{3.4}$$

where B refers to the training batch size and $Z(X_i; \Theta)$ refers to the output produced in the proposed model with Θ learnable parameters. X_i denotes the input image, and Z_i^{GT} is the ground truth of the input image X_i.

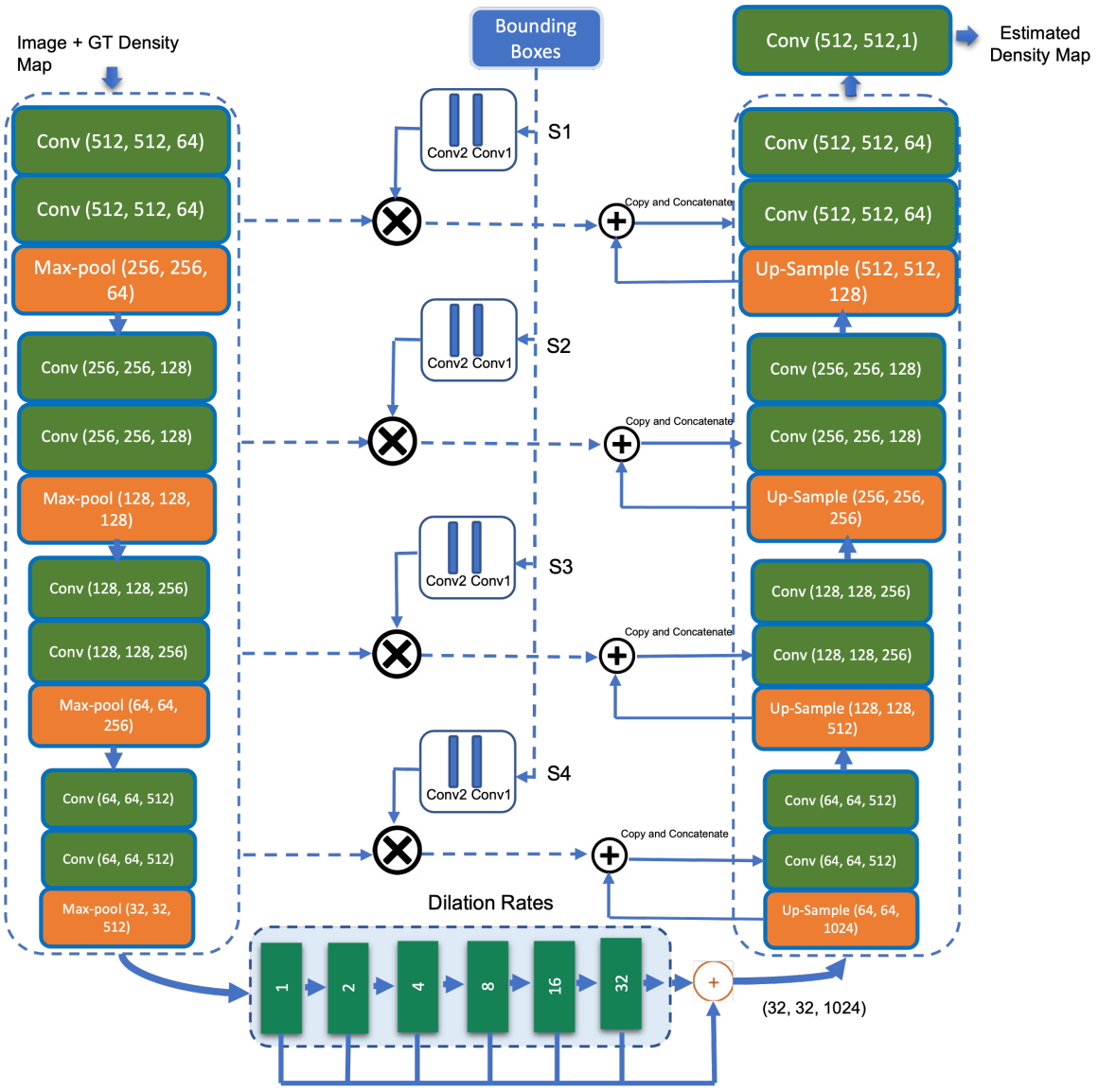


FIGURE 3.3: Block diagram of proposed dilated UNet dense regressor with attention module.

3.4 Results and Discussion

In this section, the experimental results of the proposed acne severity grading method are presented and discussed. The public dataset [143] used for assessment of the proposed model is split into 80% for training and validation (1165 images) and 20% for testing (292 images). The resolution of the facial images is fixed with 512×512 pixels. The best performance of the proposed attention-guided regressor was obtained after

TABLE 3.1: Architecture of dilated UNet layers.

Layer	Description	Size
2 Blocks-Conv	Number of filters: 64, Convolutional filter size: (3,3)	(512, 512, 64)
Max-Pool1	Max-Pool size: (2,2)	(256, 256, 64)
2 Blocks-Conv	Number of filters: 128, Convolutional filter size: (3,3)	(256, 256, 128)
Max-Pool2	Max-Pool size: (2,2)	(128, 128, 128)
2 Blocks-Conv	Number of filters: 256, Convolutional filter size: (3,3)	(128, 128, 256)
Max-Pool3	Max-Pool size: (2,2)	(64, 64, 256)
2 Blocks-Conv	Number of filters: 512, Convolutional filter size: (3,3)	(64, 64, 512)
Max-Pool4	Max-Pool size: (2,2)	(32, 32, 512)
Dilated-Conv	Number of filters: 1024, Convolutional filter size: (3,3), Six cascade blocks with dilation rates: 1, 2, 4, 8, 16, 32	(32, 32, 1024)
Up-Sampling1	Up-sampling: (2, 2)	(64, 64, 1024)
2 Blocks-Conv and Concat.	Number of filters: 512, Convolutional filter size: (3,3)	(64, 64, 512)
Up-Sampling2	Up-sampling: (2, 2)	(128, 128, 512)
2 Blocks-Conv and Concat.	Number of filters: 256, Convolutional filter size: (3,3)	(128, 128, 256)
Up-Sampling3	Up-sampling: (2, 2)	(256, 256, 256)
2 Blocks-Conv and Concat.	Number of filters: 128, Convolutional filter size: (3,3)	(256, 256, 128)
Up-Sampling4	Up-sampling: (2, 2)	(512, 512, 128)
2 Blocks-Conv and Concat.	Number of filters: 64, Convolutional filter size: (3,3)	(512, 512, 64)
Conv-DensityPrediction	Number of filters: 1, Convolutional filter size: (1,1)	(512, 512,1)

training the network for 200 epochs using the Adam optimisation method on a batch size of 4 and a learning rate of 0.0001. Data augmentation is applied to avoid over-fitting.

Table 3.2 presents the resulted in confusion matrix from the proposed model architecture, where L_0, L_1, L_2, and L_3 refer to the four severity grading levels introduced as mild, moderate, severe, and very severe labels, respectively, based on the number of inflammatory eruptions (papules, pustules) and lesions. It can be noticed that images with L_0, i.e., acne count is $\leq$5, are accurately diagnosed and graded. The remaining grading levels, L_1 (6–20), L_2 (21–50), and L_3 (>50), show that the misclassification in

the label prediction always occurs between two successive labels. For instance, L_1 is only falsely predicted as L_0. Similarly, L_2 is falsely predicted L_1, and L_3 is misclassified as L_2. This is a foreseen prediction due to the overlapping and similarity of the appearance of acne lesions with a close severity level [143].

To elaborate on the performance of the proposed method in terms of the identification of each severity level, Table 3.3 exhibits the performance evaluation in terms of precision, recall (sensitivity), specificity, accuracy and F1-Score. Unlike the default binary class confusion matrix, which considers only two classes, this problem deals with four classes producing a multi-class task. To calculate the true positive TP, true negative TN, false positive FP, and false negative FN in a multi-class task, the task is handled as a series of binary classification problems using either a One-vs-Rest (OVR) scheme or One-vs-One (OVO) scheme. Figure 3.4 illustrates the confusion matrix of the multi-class classification problem, where C_0, C_1, C_2, ..., C_n represent the classes.

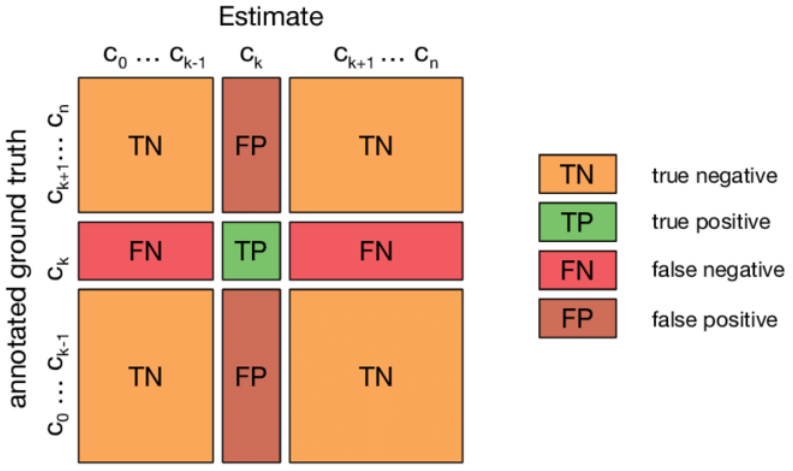


FIGURE 3.4: Confusion matrix of multi-class classification.

In OVR, also called macro-averaging, the performance metric for each class is computed individually, and then the average over classes is computed. In OVO, also called micro-averaging, the decisions for all classes (true positives, true negatives, false positives, and false negatives) are gathered into one binary-class confusion matrix. Then the evaluation metric is calculated from this binary-class confusion matrix. Thus, micro-averaging is dominated by the majority classes since the decisions for all classes are

pooled in a binary-class confusion matrix. Whereas the macro-averaging better reflects the statistics of the minority classes. In this work, the macro-averaging (OVR) scheme is adopted. The mathematical formulas used to compute the average precision (Pre), sensitivity (Sn), specificity (Sp), accuracy (Acc) and F1-Score for four lesion classes ($M = 4$) are defined as follows:

$$Pre = \frac{\sum_{n=0}^{3} Pre_{L_n}}{M} \tag{3.5}$$

$$Sn = \frac{\sum_{n=0}^{3} Sn_{L_n}}{M} \tag{3.6}$$

$$Sp = \frac{\sum_{n=0}^{3} Sp_{L_n}}{M} \tag{3.7}$$

$$Acc = \frac{\sum_{n=0}^{3} Acc_{L_n}}{M} \tag{3.8}$$

$$F1 - Score = \frac{2 \times Pre \times Sn}{Pre + Sn} \tag{3.9}$$

The last column of Table 3.3 shows the number of existing examples per each class label. In terms of precision, L_3 attains the best performance achieving a precision of 100%, followed by L_1, L_2, and L_0, respectively. The images with severity level L_0 are identified with 100% sensitivity, proving the superiority of detection in terms of true positive detection over other severity levels. Otherwise, L_1 is predicted with the lowest sensitivity, reporting only 69%. According to the true negative rate, the severity level L_3 yields the best performance with specificity 100%, whereas the severity level L_0 produces the lowest results achieving a specificity of 79%. The images with severity level L_3 (26 images) gain an accuracy of 99%, whereas the images with severity level L_1 show an accuracy of 84% (127 images). However, due to the imbalanced label distribution, the accuracy metric solely could be misleading in measuring the model performance [184].

TABLE 3.2: Confusion matrix of the proposed attention mechanism guided dilated UNet dense regressor.

	Predicted			
	L0	**L1**	**L2**	**L3**
L0	1	0	0	0
L1	0.31	0.69	0	0
L2	0	0.22	0.78	0
L3	0	0	0.15	0.85

TABLE 3.3: Performance evaluation of each class detection in the proposed attention mechanism guided dilated UNet dense regressor.

Class	**Pre**	**Sen**	**Spe**	**Acc**	**F1**	**Support**
L0	0.73	1	0.79	0.87	0.84	103
L1	0.92	0.69	0.95	0.84	0.79	127
L2	0.88	0.78	0.98	0.96	0.83	36
L3	1	0.85	1	0.99	0.92	26

Table 3.4 displays a comparison of the performance of the proposed acne grading method against methods existing in the literature. In addition to precision, sensitivity, specificity, accuracy, and F1-Score evaluation metrics; Mean Absolute Error (MAE) and Mean Square Error (MSE) are also used. These metrics can be defined as follows:

$$MAE = \frac{1}{K}\sum_{i=1}^{K}\left|C_i - C_i^{GT}\right| \tag{3.10}$$

$$MSE = \sqrt{\frac{1}{K}\sum_{i=1}^{K}\left|C_i - C_i^{GT}\right|^2} \tag{3.11}$$

where K refers to the number of test images, C_i^{GT} represents the ground truth count of acne lesions, and C_i is the estimated count of acne, which is resulted from calculating the total pixel values corresponding to acne lesions in the density map.

The number of acne lesions in an image can be counted by integrating the densities over the image region [185]. The concept of object counting from a density map was originally introduced in [185], where the integral (sum) over a region yields the number of objects in that region. A popular example of the utilisation of density maps to produce the count of objects in a region is crowd counting [165, 174, 178, 182]. Density maps in crowd counting using deep learning refer to a two-dimensional representation of the crowd distribution, where each pixel value represents the number of individuals in a local region. This representation is generated by mapping an input image of a crowd scene to a density map, where the intensity of each pixel reflects the number of individuals in that region. As explained earlier in Section 3.3.2, the ground truth density map is usually generated by applying a Gaussian kernel to the coordinates of acne lesions in the image. The summing of contributions from all acne lesions is then summed to obtain the final density map. During the model training, predicted density maps are typically created through regression, where a deep learning network is trained to predict the density map from an input image. Ultimately, the count of acne lesions from density maps can be defined using the following formula:

$$C_i = \sum_{l=1}^{L} \sum_{w=1}^{W} Z_{l,w} \tag{3.12}$$

where $Z_{l,w}$ refer to the pixel values of density map; L and W are the dimensions of density map.

For comparison purposes, results reported from state-of-the-art acne grading models summarised in Table 3.4 are broadly classified into regression-based machine learning approaches [175, 186, 187], regression-based deep learning approaches [68, 70, 72], detection-based approaches [79, 169], and label distribution learning approach [143]. In the regression-based machine learning approaches, including SIFT-Hand Crafted Features [186], HOG-Hand Crafted Features [175], and GABOR-Hand Crafted Features [187], the features SIFT, HOG, and GABOR, respectively, are extracted manually from

facial images and classified by an SVM model into four severity levels. Regression-based machine learning approaches show poor performance in all evaluation metrics. In regression-based deep learning approaches, including VGGNet [68], Inceptionv3 [70], and ResNet [72], the features are extracted automatically and fed to a fully connected neural network for classifying the severity into four levels. Contrary to the regression-based machine learning approaches, regression-based deep learning approaches achieve substantially improved performance, where ResNet [72] attains a precision of 75.81%, specificity of 91.85%, a sensitivity of 75.35%, an accuracy of 78.42% and F1-score of 75.58%. MAE and MSE metrics do not apply to regression-based methods because they use a classifier to identify the levels of acne lesion severity rather than grading based on counting the acne lesions.

TABLE 3.4: Comparison with the existing acne lesion detection and grading methods on the same dataset. NP: Not Applicable, R-ML: Regression-based Machine Learning (SVM), R-DL: Regression-based Deep Learning, D: Detection, and LD: Label Distribution.

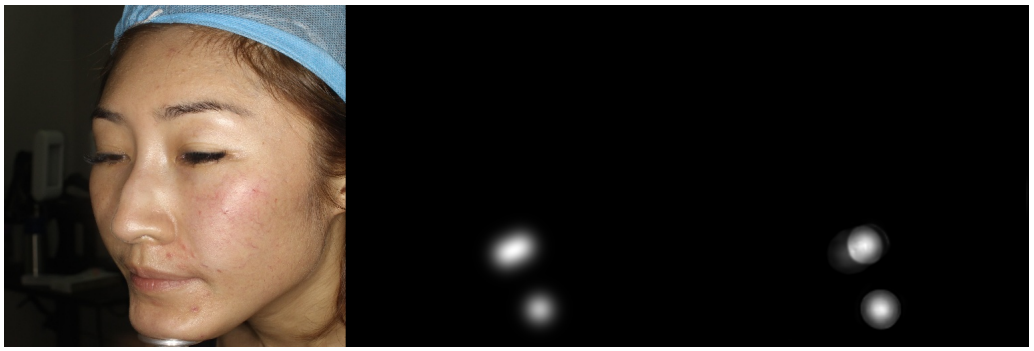
Method/Criteria	**Method Description**	**MAE**	**MSE**	**Pr**	**Sp**	**Sn**	**Acc**	**F1**
SIFT-Hand Crafted Features [186]	R-ML	NA	NA	42.59	78.44	39.09	45.89	40.77
HOG-Hand Crafted Features [175]	R-ML	NA	NA	39.1	77.91	38.1	41.3	38.59
GABOR-Hand Crafted Features [187]	R-ML	NA	NA	45.35	79.89	41.78	48.22	43.49
VGGNet [68]	R-DL	NA	NA	72.65	90.6	72.71	75.17	72.68
Inceptionv3 [70]	R-DL	NA	NA	74.26	90.95	72.77	76.44	73.51
ResNet [72]	R-DL	NA	NA	75.81	91.85	75.35	78.42	75.58
YOLOv3 [79]	D	6.69	11.35	67.01	85.96	51.68	63.7	58.35
F-RCNN [169]	D	6.7	11.51	56.91	90.32	61.01	73.97	58.89
LDL [143]	LD	2.93	5.42	84.37	**93.8**	81.52	84.11	82.92
Proposed Method	**Attention Guided Regressor**	**1.76**	**3.57**	**88.25**	93	**83**	**91.5**	**85.54**

Moreover, detection-based methods, including YOLOv3 [79] and F-RCNN [169], perform well in a sparse region where the acne lesions are not dense. However, they fail to detect when the size of the acne lesions is small and overlapped. For instance, F-RCNN [169] yields MAE of 6.7, MSE of 11.51, a precision of 56.91%, specificity of

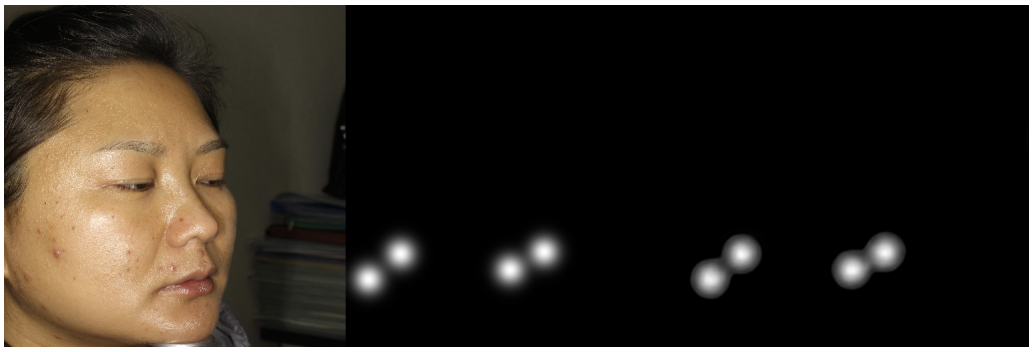
90.32%, a sensitivity of 61.01%, an accuracy of 73.97%, and F1-score of 58.89%. In the most recent acne severity grading method named LDL [143], the acne severity grading was realised following the scheme of label distribution learning (LDL) that considers the ambiguous information among levels of acne severity. The authors reported MAE, MSE, precision, specificity, sensitivity, accuracy and F1-score of 2.93, 5.42, 84.37%, 93.8%, 81.52%, 84.11%, and 82.92%, respectively. The proposed attention-guided regressor model surpasses the state-of-the-art methods in all evaluation metrics except specificity, where LDL [143] achieved better performance. The developed method shows an MAE of 1.76, MSE of 3.57, a precision of 88.35%, specificity of 93%, sensitivity of 83%, an accuracy of 91.5% and an F1-score of 85.54%. It is noteworthy to mention that the comparison depicted in Table 3.4 has been conducted using acne lesion detection and grading techniques evaluated on the same dataset utilised in the present study's developed system. This has been done with the intention of achieving a fair comparison. The performance metrics of the methods, as presented in Table 3.4, have been drawn from the referenced literature and/or reported by the authors in [143].

In terms of subjective evaluation, an example of images shown in Figure 3.5 illustrates the correct acne lesion detection and severity grading in the resulted in attention density maps using the attention mechanism guided regression model, whereas Figure 3.6 depicts the misprediction of acne lesions in the resulted in attention density maps. The Figures illustrate the attention density maps through the four levels of acne severity. These results show that the proposed model contributes to significantly estimating improved density and localisation maps. It can also be noticed the misprediction that occurred in the resulting maps is not substantial and can be tolerated. The misprediction in the density maps could be improved when training the model on a larger dataset. The presented objective and subjective performance indicate the importance of properly integrating regression and detection methods in one framework. It also reveals the significance of embedding prior knowledge into the model architecture while training. Hence, the proposed attention mechanism incorporated into regressor architecture would help to highlight salient features that are passed through the skip connections.

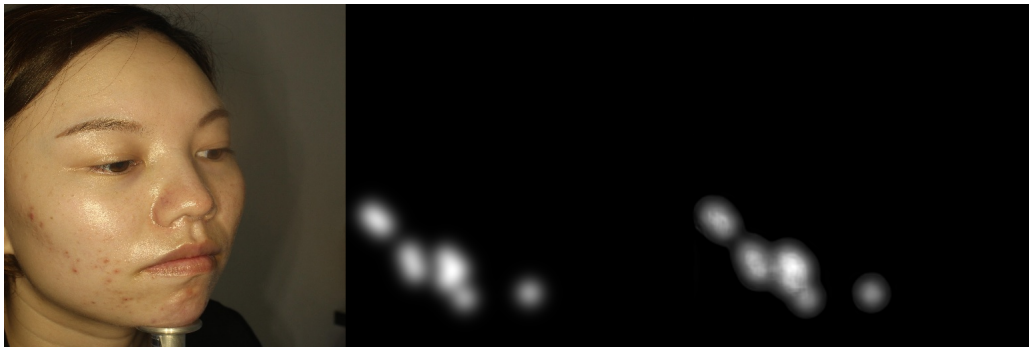
This leads us to believe that the proposed model is a viable solution when dealing with diverse object distribution in specific regions. Furthermore, the dilated convolution is shown to be a good choice, which uses sparse kernels to replace implementing several layers of the pooling and convolutional filters. In summary, this chapter presents an improved deep-learning method based on integrating regression and detection-based approaches for acne severity grading from facial images. As a result, the acne lesions are correctly counted, and the severity is accurately graded by the proposed method.



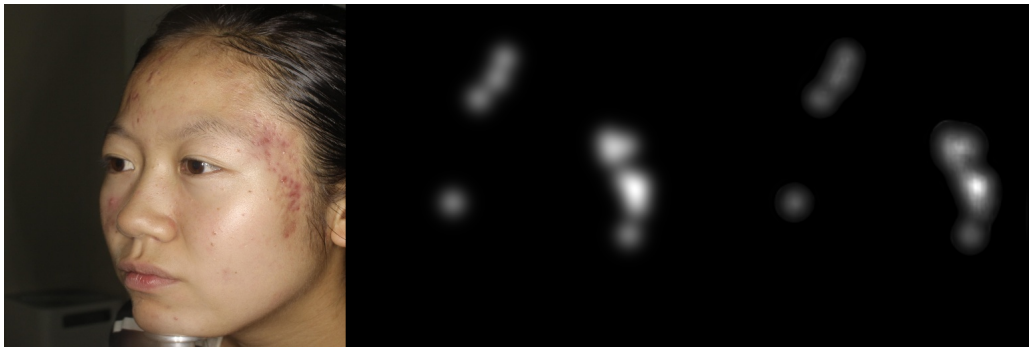
(a)



(b)

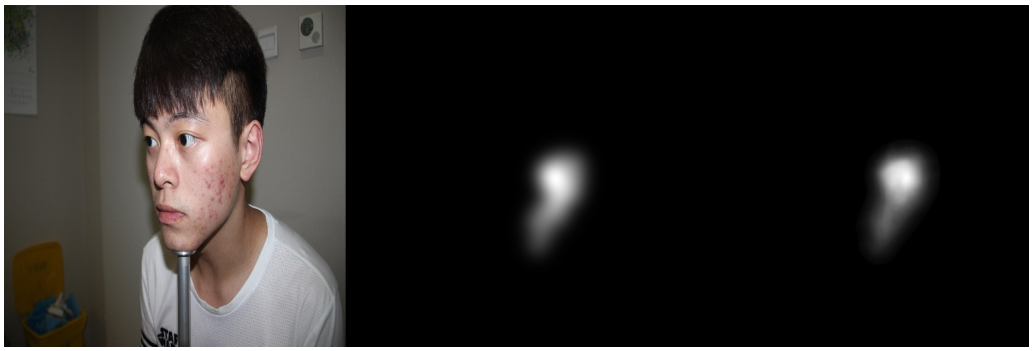


(c)

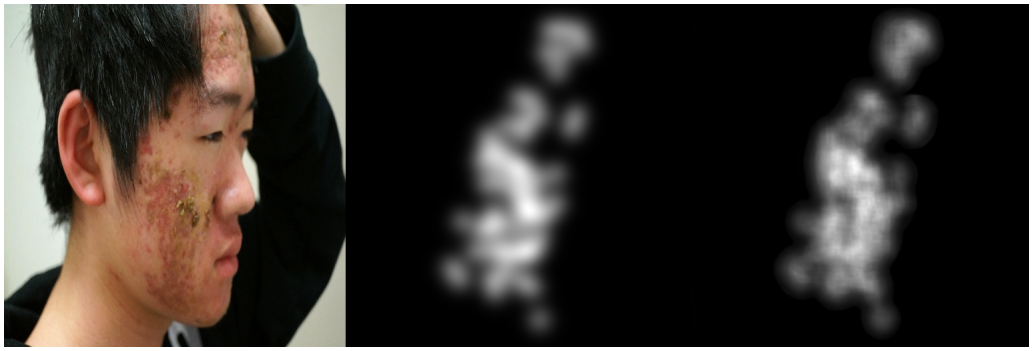


(d)

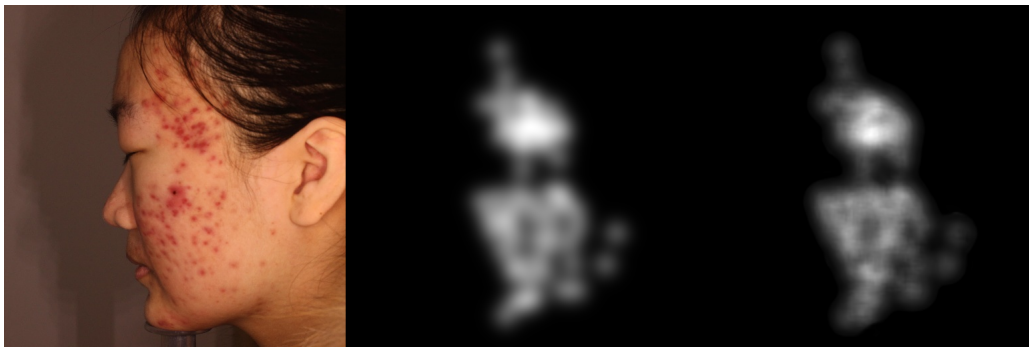
(e)



(f)

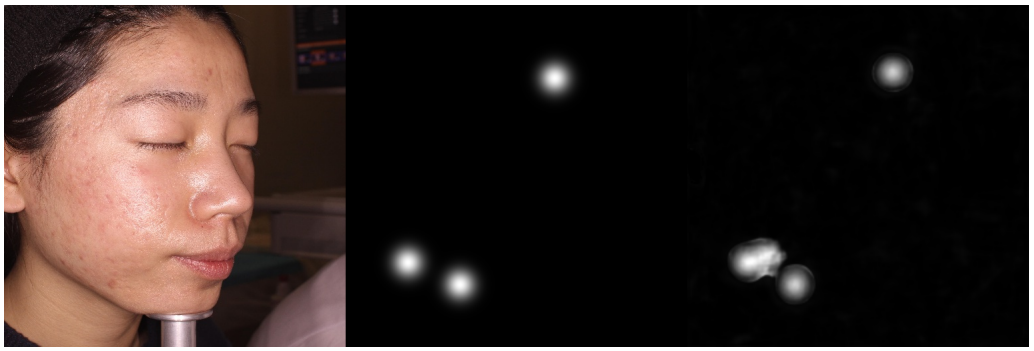


(g)

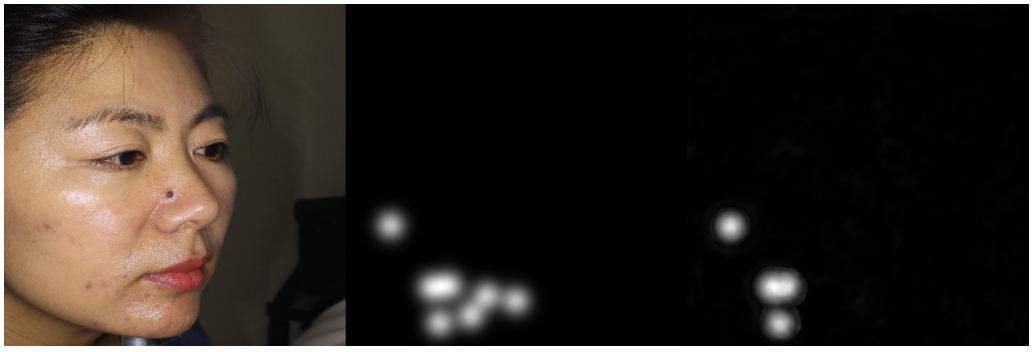


(h)

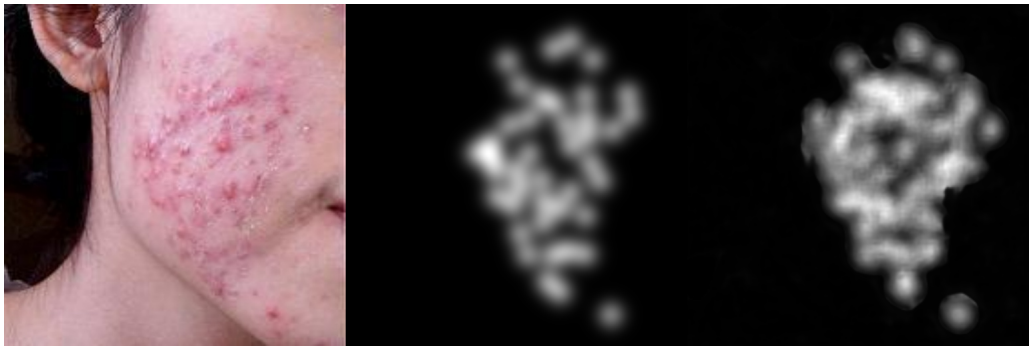
FIGURE 3.5: Image examples show correctly acne lesion detection and severity grading in the resulting attention density maps using attention mechanism guided regression model. From left to right: image, ground truth, and predicted attention density map of acne lesions. (**a**) Level 0: Example 1. (**b**) Level 0: Example 2. (**c**) Level 1: Example 1. (**d**) Level 1: Example 2. (**e**) Level 2: Example 1. (**f**) Level 2: Example 2. (**g**) Level 3: Example 1. (**h**) Level 3: Example 2.



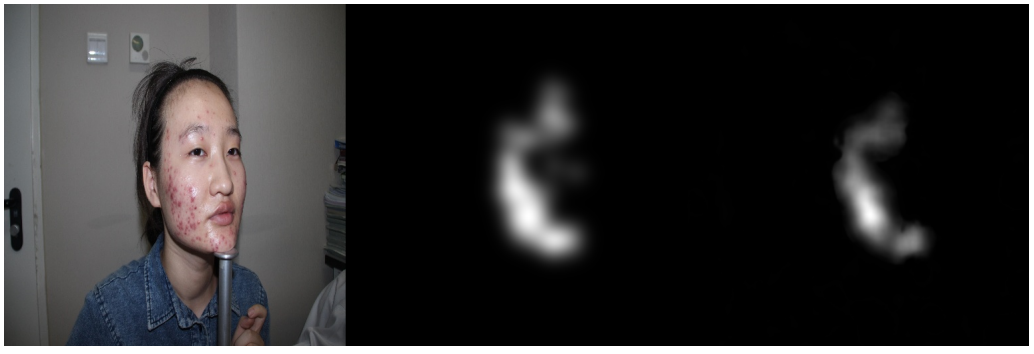
(a)



(b)



(c)



(d)

FIGURE 3.6: Image examples show misprediction of acne lesions in the resulting attention density maps. From left to right: image, ground truth, and predicted attention density map of acne lesions. (**a**) Level 0. (**b**) Level 1. (**c**) Level 2. (**d**) Level 3.

3.5 Conclusions

The work in this chapter proposed an attention mechanism integrated with dilated UNet regressor for acne counting and severity grading from two-dimensional facial images. By incorporating the attention mechanism represented by bounding boxes generated by Faster R-CNN with a density map generated by a dense regressor, following a fully supervised learning scheme, the proposed method yielded better acne grading performance than the state-of-the-art methods. Integrating bounding box information guides the proposed method to simultaneously locate the sparse and dense acne lesion regions for the density map regression task, targeting towards improving its robustness to diverse distributions of facial acne lesions.

Chapter 4

Classification of Skin Lesions Related to Melanoma From Dermoscopic Images

Unlike the regression-based model presented in the previous chapter and targeted to assess and grade the severity of acne lesions from facial images, this chapter handles another type of skin lesion, which are melanoma-related lesions. Contrary to the model developed in the previous chapter, the systems developed and investigated in this chapter are classification-based deep learning models for melanoma-related lesion diagnosis from dermoscopic images. Given the rapid development of deep learning algorithms for melanoma-related lesion diagnosis, it has become crucial to validate and benchmark these models, which is the main challenge of the work presented in this chapter. The outcome of the research reported in this chapter was published in Alzahrani et al. [188]. The contribution of the first author to work presented in this chapter is the conceptualisation of the idea, proposing the methodology, development of the models, data and results analysis and writing up. The labelled dataset used for training and testing was collected from the openly available International Skin Imaging Collaboration (ISIC 2017) dataset [126].

4.1 Introduction

The initial stage in melanoma diagnosis is usually a visual assessment of the skin lesions. In comparison to inspection with the naked eye, dermatoscopy is one of the dermatologists' most popular imaging procedures, and a frequently used diagnostic tool that enhances and improves the diagnosis of malignant and benign pigmented skin lesions [28].

However, manual review by dermatologists is a time-consuming, controversial, and error-prone task. The number of required dermatologists comparing the size of the population in the United States, Australia, and the UK is considerably low [189–191]. In the USA, the required number of dermatologists should be more than 4 per 100000 individuals, which is the number that has been suggested to care for a population adequately. However, it is currently estimated at 3.4 per 100000 individuals. Similarly, there are just 550 practising dermatologists in Australia, which is almost 15 % less than what is required to meet the needs of the population [190]. In the UK, the Royal College of Physicians (RCP) [192] recommends one full-time equivalent (FTE) consultant per 62,500 population. The RCP recommends 989 FTE consultant dermatologists. The British Association of Dermatologists (BAD) [193] found there were 813 dermatology specialists in the UK. Compared to the RCP's recommendations, the BAD show a shortfall in the region of 250 consultants [191]. Hence, melanoma patients may not be aware of the severity of their disease if they do not have the inspection by skilled specialists in the early stage of the disease and thus miss the ideal time to treat their conditions.

These obstacles encourage and inspire researchers to create automated melanoma diagnosis methods using computer-aided diagnosis (CAD) systems. For non-experienced dermatologists, the CAD tool could provide a user-friendly environment used as a second opinion in melanoma cancer diagnosis [194, 195]. A large volume of skin images has been collected in recent years, and sophisticated deep learning-based models [196] have been successfully trained to perform automatic analysis of these skin images due to the

industrial advancement of both computer hardware represented by graphics card capabilities and software technologies. These breakthroughs prompted expectations that automated diagnostic tools will be available in the near future to examine and diagnose all types of skin lesions without the requirement for human experience [126].

Many automated melanoma diagnosis systems based on deep learning techniques, especially deep convolutional neural networks (CNNs), have been recently developed. The new methods have significantly advanced the state-of-the-art in skin lesion analysis. The CNN can automatically extract and learn high-level features, increasing the robustness of melanoma images' inter-and intra-class variability [197, 198]. With the rapid increase in the number of automatic recognition of melanoma from dermoscopy images using CNNs, comparing results among pieces of work and evaluation has become an awkward task. This limitation is due to methodological constraints and the absence of some of the standard metrics used to evaluate the performance of the models in terms of sensitivity, specificity, accuracy, etc. To overcome these limitations, the deep learning models applied for melanoma diagnosis have been assessed and benchmarked by considering similar methodological constraints, similar experimental settings and parameters setup, and similar evaluation criteria for all the deep learning models used in this study. Due to the existence of trade-offs and conflict among performance evaluation criteria during the evaluation process, the benchmarking of DL models is dealt with as a multiple criteria problem [199]. Accordingly, multi-criteria decision-making schemes (MCDM) can be exploited to benchmark the convolutional neural network models used for melanoma diagnosis.

Multi-criteria decision-making methods (MCDM) are an application of decision theory that handles multi-objective choice. It's a strategy for assessing and comparing multiple solutions (alternatives) considering competing criteria. It is a widely used decision-making approach in the field of operational research that deals with several criteria to find an optimal solution for decision-makers. MCDM techniques find the optimal selection by ranking the performance of the alternatives where the highest rank is assigned to the best feasible alternative (solution) [199–201]. Two key problems

could be arisen during the evaluation and benchmarking of deep convolutional neural network models for melanoma detection. First, what are the suitable criteria for the evaluation? And second, what is the proper benchmarking approach for selecting the optimal model considering the provided criteria? Thus, the motivation of this work is to present a framework for evaluating and benchmarking multiple deep-learning models for melanoma detection using various evaluation criteria.

In light of the concerns mentioned above and given the rapid development of deep learning algorithms for melanoma diagnosis, it is crucial to validate and benchmark these models, which is the main challenge of this work. This research direction aims to conduct a comprehensive evaluation and benchmark of convolutional neural networks for melanoma diagnosis. The benchmarking is accomplished by prioritising the convolutional network architectures and then selecting the optimal architecture given specific criteria.

4.2 Materials and Methods

4.2.1 Materials

To carry out our experiments, dermoscopic images were collected from the openly available International Skin Imaging Collaboration (ISIC 2017) dataset [126]. Melanoma, seborrheic keratosis, and nevus, shown in Figure 4.1, are the three types of lesions represented in the dataset. Melanoma is a cancerous skin tumour with a high mortality rate. Seborrheic keratosis and nevus, the other two types of lesions, are benign skin tumours formed from different cells. Despite the ISIC Challenge 2017 included three subtasks with annotations for three classes (nevus, seborrheic keratosis, and melanoma), the melanoma subtask versus the remaining classes is only considered, producing a two-class classification task. The ISIC (2017) dataset comprises 2000 training images and

600 test images. In the training set, there are 374 melanoma images and 1626 non-melanoma. The test set contains 117 melanoma images and 483 non-melanoma images. In total, both training and test data comprise 491 melanoma images and 2109 non-melanoma images. The percentage of melanoma images in the dataset is 19%. This ratio shows a highly imbalanced data distribution between the two classes. The presented study does not target to develop a new method for melanoma diagnosis competing with other methods in which particular strategies are designed to remedy and alleviate the effect of imbalanced data. Instead, this study aims to evaluate and benchmark the existing CNNs architectures considering multiple conflicting criteria. The condition of benchmarking in this study is set for balanced data. Thus, to maintain the balance of class distribution, all the melanoma images (491) in the dataset are collected, whereas only the first 500 non-melanoma images are gathered, producing 991 dermoscopic images in total. The data has been split into five folds for training and testing. In each of the five training cycles, four folds are used for training and the hold-out set is used for testing the network performance. Thus, in each training process, this will generate 393 images (melanoma) and 400 images (non-melanoma) for training, and 98 images (melanoma) and 100 images (non-melanoma) for testing.

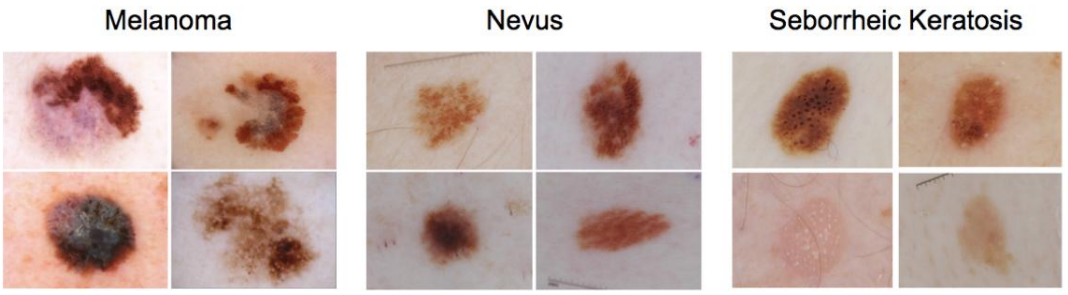


FIGURE 4.1: Example of images used to conduct this study. Both nevus and seborrhoeic keratosis are classified as non-melanoma in the conducted experiments

.

4.2.2 Methods

The developed evaluation and the benchmarking system illustrated in Figure 4.2 comprises five main stages, including data preparation, designing of CNN models, training of CNN models, evaluation criteria establishment, and benchmarking of CNN models using MCDM. In the first and second phases of the proposed framework, depicted as red and orange blocks in Figure 4.2, the data is prepared, and deep convolutional neural networks are implemented (different versions of a specific CNN architecture are considered, for instance; VGG16, VGg19). In the third phase, depicted as a grey block, the CNN models are trained. The key evaluation criteria are identified and measured by evaluating the trained models on test data. In the final phases, shown as blue and green blocks, MCDM methods are employed to prioritise the alternatives (i.e., CNN models). The blue block shows the construction of the decision matrix (models as rows and criteria as columns), then applies the entropy method to calculate and generate the weights of criteria. Finally, the MCDM methods (PROMETHEE and VIKOR) are exploited to rank CNN models and report the optimal CNN architecture considering the provided decision matrix and the weights of criteria. Although PROMETHEE and VIKOR are different statistical methods, the input data of these methods is the same, which are the weights of criteria and the decision matrix. These methods are independent; therefore, they were applied to the given input data separately. In this section, each phase of the proposed framework is described as follows:

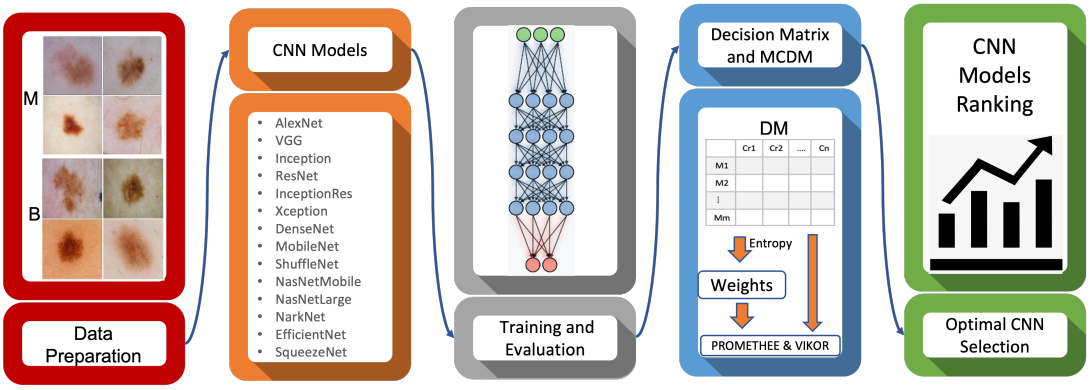


FIGURE 4.2: The block diagram of the proposed framework used to benchmark CNN models for melanoma diagnosis. M refers to malignant (melanoma) and B refers to benign (non-melanoma).

Pre-trained convolutional neural network models (CNNs)

The key CNN baseline architectures that have been applied in this study are summarised below:

- ***AlexNet:*** In 2012, AlexNet [48] substantially surpassed all previous classification methods, winning the ImageNet Large Scale Visual Recognition Competition (ILSVRC) by reducing top-5 error from 26% to 15.33%. The network's design was similar to the LeNet network developed by Yann LeCun et al. [86], but it was deeper, with more filters per layer and layered convolutional layers. 11×11, 5×5, 3×3 convolutions filters, max pooling, dropout, data augmentation, ReLU activations, and SGD with momentum were all part of it. After each convolutional layer, it added ReLU activations. AlexNet was trained using two Nvidia Geforce GTX 580 GPUs for six days, which is why their network is divided into two pipelines.

- ***VGG16,19:*** Simonyan and Zisserman presented the VGG architecture in 2014 [68]. It is a straightforward design, with only blocks made up of an incremental number of convolution layers and 3x3 filters. Furthermore, max-pooling blocks follow convolution blocks to reduce the size of the activation maps obtained. Finally, a classification block is employed, consisting of two dense layers and a final output layer. The numbers 16 and 19 refer to how many weighted layers each network includes. On the other hand, this network has a couple of drawbacks: it takes too long to learn and has a lot of parameters.

- ***InceptionV1,V3:*** Goole implemented inception building blocks in GoogLeNet (Inceptionv1) [67]. These blocks function well together and result in a model that is easy to generalise. GoogLeNet is made up of nine Inception modules that are stacked one on top of the other. There are a total of 27 layers, 5 of which are pooling layers. The total number of layers used in the network design is

about 100. New revisions of the model appeared as the model was updated regularly. Inception-v2 and Inception-v3 [70] were released within a short time gap in 2015. Except for a few features, Inception-v2 integrates all of GoogLeNet's features. Filter banks were increased in width in Inception-v2 to eliminate the "representational bottleneck". All of the changes from Inception-v2 were included in Inception-v3. Furthermore, Inception-v3 were undergone additional changes, such as the use of a higher resolution input and the use of the RMSProp optimiser, which significantly reduced the cost function.

- ***InceptionResNetV2:*** Inception V4 was launched in 2016 by Google researchers in conjunction with Inception-ResNet. By implementing Inception -V4, the main goal of this network architecture was to reduce the complexity of the Inception V3 model, which provided state-of-the-art accuracy on the ILSVRC2015 challenge. This architecture also investigates the use of residual networks on the Inception model [69].

- ***ResNet18,50,101:*** The ResNet architecture, founded by He et al. in 2015 [72], was a major turning point in the introduction of an extraordinary form of architecture focused on "modules" or "networks within networks". The principle of residual connections was first implemented in these networks. ResNet comes in various sizes and numbers of layers, like ResNet18, RerNet50, and RerNet101, but the most common is ResNet50, which has 50 layers with weights. Despite having many more layers than the VGG, ResNet50 needs nearly five times less memory. This is because, instead of dense layers, this network uses a layer called GlobalAveragePooling in the classification stage, which transforms the 2D feature maps of the last layer in the feature extraction stage into an n-classes vector that is used to measure the likelihood of belonging to each class.

- ***DenseNet201:*** DenseNet [71] is very similar to ResNet, but there are a few key differences. DenseNet concatenates the output of the previous layer with

the output of the next layer. At the same time, ResNet follows an additive approach that combines the previous layer (identity) with the next layer. DenseNet model was founded mainly to address the vanishing gradient's impact on high-level neural networks' layers. Using the composite function operation, the previous layer's output becomes the second layer's input. Convolution, pooling, batch normalisation, and non-linear activation layers form this composite process. DenseNet comes in a variety of types, including DenseNet-121, DenseNet-169, and DenseNet-201. The numbers represent the number of the neural network's layers.

- ***Xception:*** Xception [74] is an extension of the Inception architecture that uses depthwise separable convolutions to replace the regular Inception modules. The mapping of cross-channel and spatial correlations in the feature maps of convolutional neural networks can be fully decoupled in this network. The authors called their proposed architecture Xception, which stands for "Extreme Inception," since this hypothesis is a stronger version of the hypothesis that underlies the Inception architecture. In a nutshell, the Xception architecture is a depthwise separable convolution layer stack with residual connections. This makes it very simple to establish and change the architecture.

- ***MobileNet:*** MobileNet [73] is a convolutional neural network designed for mobile and embedded vision uses. They are based on a streamlined architecture that builds lightweight deep neural networks with low latency for mobile and embedded devices using depthwise separable convolutions. The Width Multiplier and Resolution Multiplier parameters are added to make it easier to tune MobileNet. The depthwise convolution in MobileNets applies a single filter to each input channel. After that, the pointwise convolution applies a 1×1 convolution to combine the depthwise convolution's outputs. A separate layer for filtering and a separate layer for combining are used in depthwise separable convolution. This factorisation has the effect of reducing computation and model size drastically.

- ***NASNetMobile and NASNetLarge:*** Google Brain built Neural Architecture Search (NASNet) [76]. The authors suggested that an architectural building block be detected on a small dataset and then transferred to a larger dataset. They generally look for the best convolutional layer or cell on a small dataset first, then stack together more copies of this cell to extend to the larger dataset. Besides, a new regularisation technique called ScheduledDropPath is proposed, which significantly enhances the generalisation of NASNet models. With a smaller model size and lower complexity, the NASNet method achieved state-of-the-art results. While the overall architecture of NASNet is predefined, the blocks or cells are not. Alternatively, a reinforcement learning search technique is used to find them. The authors developed different versions of NASNets with different computational requirements. The larger model, NASNetlarge, is a convolutional neural network trained on over a million images from the ImageNet database, while the smaller model, NASNetMobile, is optimised for mobile devices.

- ***ShuffleNet:*** ShuffleNet [75] is a convolutional neural network optimised for mobile devices with minimal processing capacity developed by Megvii Inc (Face++). The network architecture design considers two new operations to lower computation costs while retaining accuracy: pointwise group convolution and channel shuffle. It specialises in common mobile platforms such as drones, robots, and smartphones and aims for the best accuracy in minimal computational resources.

- ***DarkNet19,53:*** The backbone of YOLOv2 is a convolutional neural network called Darknet-19 [77]. It generally employs 3×3 filters and twice the number of channels after each pooling phase, similar to VGG models. It leverages global average pooling to produce predictions and 1×1 filters to compress the feature representation among 3×3 convolutions, identical to the work on Network in Network (NIN). Batch normalisation is a technique for stabilising training and accelerating convergence. Darknet-53 [79], on the other hand, is a convolutional neural network that serves as the backbone for the YOLOv3 object

detection method. The utilisation of residual connections and more layers is an enhancement over its predecessor, Darknet-19.

- ***EfficientNetB0:*** EfficientNetB0 [78] is a convolutional neural network that scales depth, width, and resolution dimensions using a compound coefficient. Unlike the traditional methodology, which arbitrarily scales network dimensions, the EfficientNetB0 scaling strategy scales network dimensions with a set of predetermined scaling coefficients. According to the compound scaling approach, if the input image is larger, the network needs more layers and channels to widen the receptive field and catch more fine-grained patterns on the larger image. In addition to squeeze-and-excitation blocks [202], the base of EfficientNet is built on MobileNetV2's inverted bottleneck residual blocks [73].

- ***SqueezeNet:*** DeepScale, UC Berkeley, and Stanford University collaborated to develop SqueezeNet [80]. With 50x fewer parameters, SqueezeNet reaches AlexNet-level accuracy on ImageNet. Additionally, the authors were able to compress SqueezeNet to less than 0.5MB using model compression approaches (510x smaller than AlexNet). Smaller Convolutional Neural Networks (CNNs) require less communication across servers during distributed training and less bandwidth. They are also more feasible to be deployed on FPGAs and hardware with restricted computational resources and limited memory.

Benchmarking Criteria

This section presents an elaboration on the criteria taken into consideration in this study. The choice of criteria in MCDM methods is highly dependent on the decision-making context, and the problem handled. As the problem targeting to deal with here is a classification task, the presented study has established the most popular measurements typically used for classifiers' evaluation as criteria. The performance of each CNN model was evaluated in this stage using ten evaluation metrics. The test accuracy, F1-score, sensitivity, specificity, precision, false-positive rate and false-negative rate, Matthews

correlation coefficient (MCC), classification error, and network complexity have been utilised to evaluate each of the models targeted for study in this research.

- ***Accuracy:*** this metric measures how close the predicted value is to the actual data values. It can be defined using the following formula:

$$Accuracy\ (Acc) = \frac{tp + tn}{tp + tn + fp + fn} \tag{4.1}$$

tp: True Positive, tn: True Negative, fp: False Positive, fn: False Negative

- ***Classification Error:*** refers to the number of samples incorrectly classified (false positives and false negatives). It can be defined as follows:

$$Classification\ Error\ (Err) = 1 - Acc \tag{4.2}$$

- ***Precision:*** the precision metric tests the ability of the classifier to reject irrelevant samples. The formula of this metric can be defined as follows:

$$Precision\ (Pre) = \frac{tp}{tp + fp} \tag{4.3}$$

- ***Sensitivity:*** the sensitivity metric measures the proportion of the correctly detected relevant samples. It can be represented as follows:

$$Sensitivity\ (Sn) = \frac{tp}{tp + fn} \tag{4.4}$$

- ***F1-Score:*** the F1-score can be obtained by the weighted average of sensitivity (recall) and precision, where the relative contribution of both recall and precision to the F1-score are equal. The F1-Score can be defined as follows:

$$F_1\ Score = \frac{2(Precision \times Recall)}{Precision + Recall} \tag{4.5}$$

where $Recall$ = $Sensitivity$

- ***Specificity:*** it describes the ability of the classifier to detect the true negative rate. The formula of specificity can be defined using the following equation:

$$Specificity\ (Sp) = \frac{tn}{tn + fp} \tag{4.6}$$

- ***False-Positive Rate (FPR):*** is the proportion of negative examples wrongly categorised as positive. This metric is also known as the miss rate and is represented as:

$$False - Positive\ Rate\ (FPR) = \frac{fp}{fp + tn} \tag{4.7}$$

- ***False-Negative rate (FNR):*** is the proportion of positive examples wrongly categorised as negative. This metric is also known as the fall-out rate. This evaluation criterion is introduced as follows:

$$False - Negative\ Rate\ (FNR) = \frac{fn}{fn + tp} \tag{4.8}$$

- ***Matthews Correlation Coefficient (MCC):*** The MCC is a correlation coefficient that yields a value between -1 and $+1$ for actual and estimated binary classifications. A coefficient of $+1$ shows ideal prediction, 0 shows random prediction, and -1 indicates complete disagreement between predictions and the ground truth. MCC can be defined as:

$$MCC = \frac{(t_p \times t_n - f_p \times f_n)}{\sqrt{(t_p + f_p)\,(t_p + f_n)\,(t_n + f_p)\,(t_n + f_n)}} \tag{4.9}$$

- ***CNN Complexity:*** refers to the number of parameters existing in the pre-trained CNN.

Multi-Criteria Decision Making (MCDM)

Multi-criteria decision-making typically involves six phases: (i) problem formulation, (ii) identification of requirements, (iii) goal setting, (iv) identification of alternatives, (v) development of criteria, and (vi) identification and application of decision-making techniques. This process can be carried out using various mathematical procedures chosen based on the problem at hand, and the level of complexity ascribed to the decision-making process [203, 204]. This study has formulated the CNN models benchmarking as the research goal, considering nineteen CNNs as alternatives and ten criteria. For decision-making, Preference Ranking Organization Method for Enrichment Evaluation (PROMETHEE) [205], an MCDM method, is adopted to generate the ranking list and to produce the optimal model selection using the criteria's weights computed by the Entropy method. For validating the optimal model selection, another MCDM method called VlseKriterijumska Optimizacija I Kompromisno Resenje (VIKOR) in Serbian [206], which means multi-criteria optimisation and compromise solution, is also applied. This section describes the MCDM methods exploited to rank the CNN models and select the optimal model given the criteria mentioned earlier using data in the presented case study.

- ***Entropy:*** This method computes relative weights by objectively interpreting the relative intensities of the criteria significance based on data discrimination [207]. MDCM's generated decision matrix DM is defined by m alternatives (nineteen CNN models) and k criteria (ten criteria), which are represented as follows:

$$DM = [x_{ij}]_{m \times k} \tag{4.10}$$

From the constructed decision matrix DM, the procedure of entropy weighting method described in [207] is followed to measure the weights w_j. x_{ij} refers to each entry in the DM, where $i = 1, \ldots, m$, $j = 1, \ldots, k$. The steps of entropy

weighting method [207] is described as follows:

Step1: Normalising the decision matrix using the following equation:

$$p_{ij} = \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}}, (1 \leq i \leq m, 1 \leq j \leq k) \tag{4.11}$$

Step2: Measure the entropy of criteria:

$$e_j = -g \sum_{j=1}^{k} p_{ij} \ln p_{ij}, (g = 1/\ln m, 1 \leq i \leq m, 1 \leq j \leq k) \tag{4.12}$$

.

Step3: Determine the inherent contrast intensity:

$$d_i = 1 - e_j, (1 \leq j \leq k) \tag{4.13}$$

Step4: The entropy weights of criteria are then defined as follows:

$$w_j = d_j / \sum_{j=1}^{k} d_j, (1 \leq j \leq k) \tag{4.14}$$

- ***PROMETHEE:*** The PROMETHEE is an outranking approach for ranking and selecting a finite collection of alternatives based on often competing criteria. Compared to other multi-criteria analysis methods, PROMETHEE II is an uncomplicated complete (not partial) ranking method in terms of conception and application. The stepwise procedure of PROMETHEE II can be defined as follows, giving the provided decision matrix and the weights of criteria:

Step 1: Determining of deviations based on pairwise comparisons as follows:

$$d_j(a,b) = g_j(a) - g_j(b) \tag{4.15}$$

where $d_j(a,b)$ refers to the difference between the evaluations of a and b on each criterion.

Step 2: Preference function application:

$$P_j(a,b) = F_j\left[d_j(a,b)\right] \quad j = 1, \ldots, k \tag{4.16}$$

Where $P_j(a,b)$ denotes the preference of alternative a with regard to alternative b on each criterion, as a function of $d_j(a,b)$.

Step 3: Calculating an overall or global preference index using the following formula:

$$\pi(a,b) = \sum_{j=1}^{k} P_j(a,b) w_j \tag{4.17}$$

Where $\pi(a,b)$ of a over b represents the weighted sum $p(a,b)$ for each criterion, and w_j is the weight w_j related to the j th criterion.

Step 4: Calculating the partial ranking PROMETHEE I (outranking flows) using the following equations:

$$\phi^+(a) = \frac{1}{m-1} \sum_{b=1}^{m} \pi(a,b) \tag{4.18}$$

$$\phi^-(a) = \frac{1}{m-1} \sum_{b=1}^{m} \pi(b,a) \tag{4.19}$$

Where $\phi^+(a)$ and $\phi^-(a)$ represent the positive outranking flow and negative outranking flow for each alternative, respectively.

Step 5: Calculating the complete ranking PROMETHEE II (outranking flows) using the following equations:

$$\phi(a) = \phi^{+}(a) - \phi^{-}(a) \tag{4.20}$$

Where $\phi(a)$ represents the outranking flow for each alternative.

- ***VIKOR:*** The VIKOR approach [206] was initially developed to optimise complex systems that involve various parameters. Using the predefined weights, the VIKOR provides a compromise ranking list and suggests a compromise solution. VIKOR creates a multi-criteria rating index based on a specific "closeness" metric to the "ideal" solutions [206]. The VIKOR methodology's compromise ranking algorithm can be described as follows, giving the provided decision matrix and the weights of criteria:

  Step1: Determining the best value as ${x_j}^*$ and the worst value as ${x_j}^-$ of the criteria as $j = 1, 2, ..., k$. This also leads to configuring the criteria as beneficial and non-beneficial values. The beneficial attributes require to be maximised while the non-beneficial need to be minimised, which are identified as follows:

  Rule1: Best value for beneficial criteria is ${x_j}^* = max x_{ij}$, and for non-beneficial is ${x_j}^* = min x_{ij}$,

  Rule2: Worst value for beneficial criteria is ${x_j}^- = min x_{ij}$, and for non-beneficial is ${x_j}^- = max x_{ij}$.

  Step2: Determining the values of S_i and R_i, where $i = 1, 2, ..., m$ using the following equations:

$$\begin{aligned} S_i &= \sum_{j=1}^{k} w_j \left(x_j^* - x_{ij}\right) / \left(x_j^* - x_j^-\right), \\ R_i &= \max_j w_j \left(x_j^* - x_{ij}\right) / \left(x_j^* - x_j^-\right), \end{aligned} \tag{4.21}$$

where w_j are the weights of criteria computed using the entropy method.

Step3: Determining the values of S^* and R^* as follows:

$$\begin{aligned} S^* &= \min_i S_i, R^* = \min_i R_i, \\ S^- &= \max_i S_i, R^- = \max_i R_i \end{aligned} \tag{4.22}$$

Step4: Determining the values of Q_i; where $i = 1, 2, \ldots, m$ and v is defined as the weight of the scheme of "the majority of criteria" using the following equation:

$$Q_i = v\left(S_i - S^*\right) / \left(S^- - S^*\right) + (1 - v)\left(R_i - R^*\right) / \left(R^- - R^*\right) \tag{4.23}$$

Step5: Ranking the alternatives by sorting the values of Q_i in ascending order.

4.3 Experimental Results and Discussion

4.3.1 Experimental Setup and Training

During the experimental process, nineteen CNN models pre-trained on ImageNet dataset [208] were modified and re-trained using transfer learning and fine-tuning strategies to classify the skin lesion into two classes: cancerous (melanoma) or non-cancerous (non-melanoma). The characteristics of the CNN architectures in terms of the number of total layers, the number of learnable layers, the size of CNN, the size of the input image, and the number of parameters in each network architecture are described in Table 4.1. In the training of models, binary cross-entropy was preferred as a cost function, and the stochastic gradient descent with momentum (SGDM) optimiser was used to minimise the cost function. The softmax activation function was used in the output layer of the models. Each model was trained through six epochs, and the training was repeated a

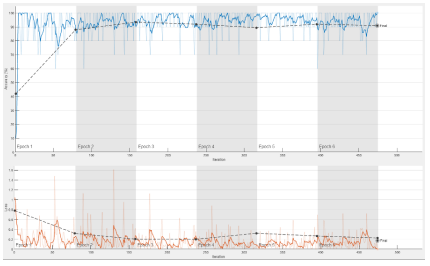
total of five times. The batch size is set to 10, providing 79 iterations per epoch and 474 iterations for six epochs. The learning rate value was set to 0.0003 and momentum to 0.9.

The learning curves (i.e. training and validation curves) of nineteen CNN models are presented in Figure 4.3. These curves provide insight into how the model is learning and how well it is performing during training. The training curve shows the accuracy or loss of the model on the training data over the course of training epochs. The goal is for the training accuracy/loss to increase/decrease over time, indicating that the model is learning from the training data and improving its predictions. If the training accuracy/loss starts to decrease/increase, it may indicate overfitting, where the model has memorised the training data and is not generalising well to new data. The validation curve shows the accuracy or loss of the model on a validation set, which is a separate set of data that is used to evaluate the model's performance during training. The validation accuracy/loss provides a measure of the model's ability to generalise to new, unseen data. It is important to monitor the validation accuracy/loss during training to ensure that the model is not overfitting the training data. This information can help diagnose issues such as overfitting, underfitting, or convergence problems, and make adjustments to the model or training process as needed.

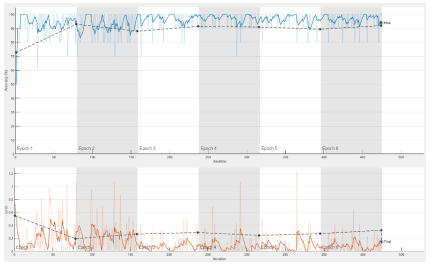
To provide fair performance evaluation and benchmark among the nineteen models, a fixed number of epochs for all models is opted to use. Figure 4.3 shows that all the models stopped training at the same endpoint, and the trained models have been deployed from this endpoint to conduct the testing phase. It is aimed to compare the performance of the networks under the same constraints and conditions. So, choosing the optimal number of epochs to train a particular model has not been considered. Considering learning the models under the same conditions, if one model encountered overfitting and subsequently failed to achieve good accuracy on the unseen test set, whereas another model has not undergone overfitting, the later model is preferred over the first model. However, in Figure 4.3, it can be noticed that the training and validation curves show a steady learning behaviour, and there is no indication of overfitting.

In order to prevent potential overfitting during the training, the online data augmentation was applied by using various image transformation methods such as vertical and horizontal flipping, random translation in the range of $[-30, 30]$, and random scaling in the range of $[0.9, 1.1]$. In most of the CNN models, the last layer is the learnable weights of fully connected layers. Thus, to apply the transfer learning and fine-tune the network using the provided data, these completely connected layers are replaced with a new fully-connected layer comprising two neurons adhering to the two classes in the presented study. Instead of fully connected layers, the last learnable layer in some networks, such as SqueezeNet, is a 1×1 convolutional layer. In this scenario, the old convolutional layer is replaced by a new convolutional layer with the same number of filters as classes.

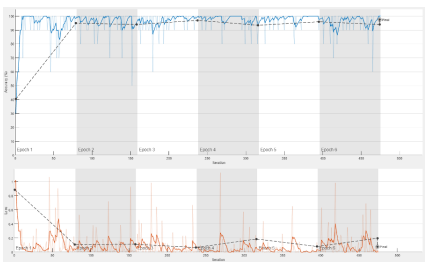
(a)



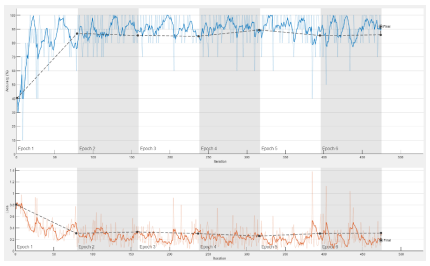
(b)



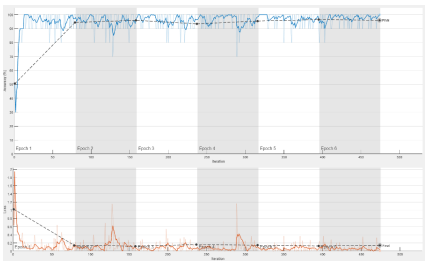
(c)



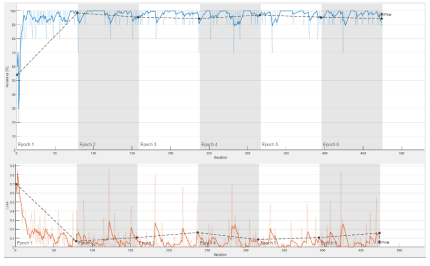
(d)



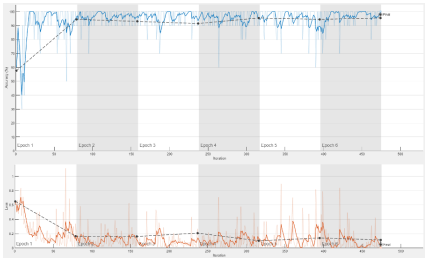
(e)



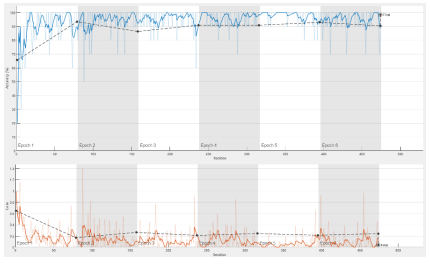
(f)



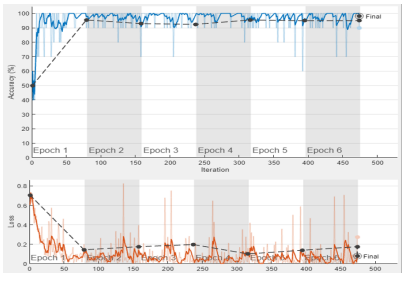
(g)



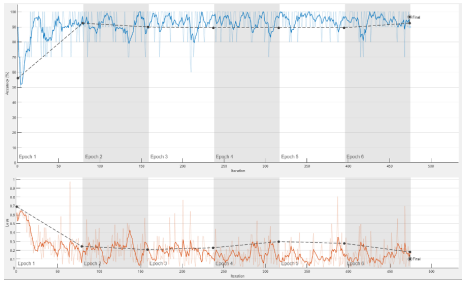
(h)



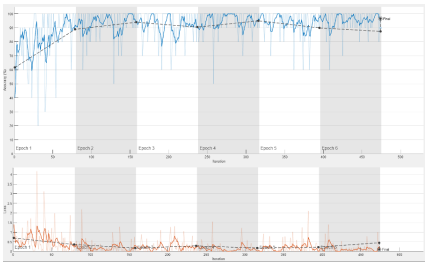
(i)


Accuracy (%)
Final
Epoch 1
Epoch 2
Epoch 3
Epoch 4
Epoch 5
Epoch 6
Iteration
Loss
Epoch
Final

(j)

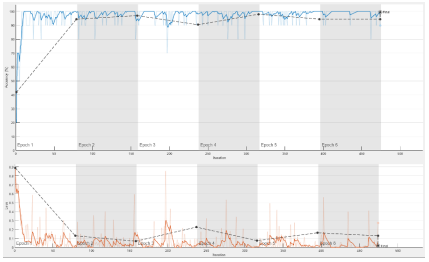


(k)

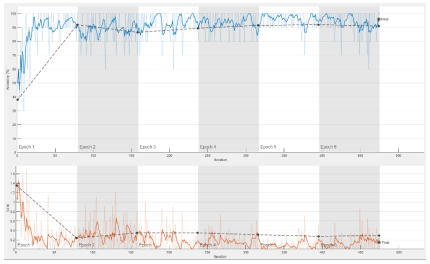


(l)

(m)

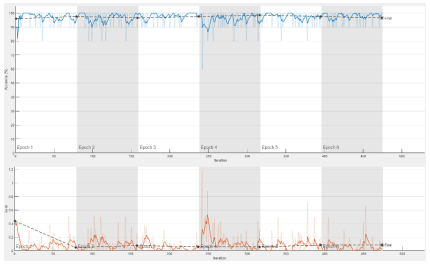


(n)

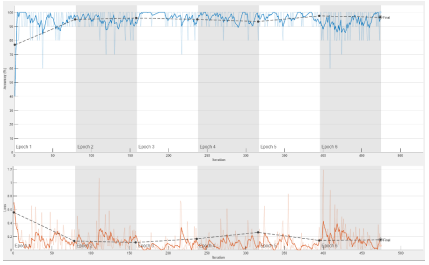


(o)

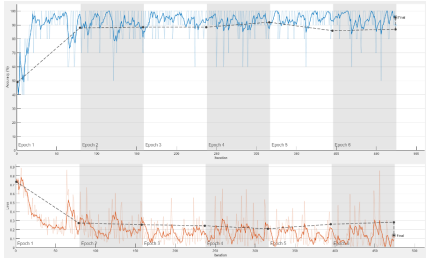
(p)



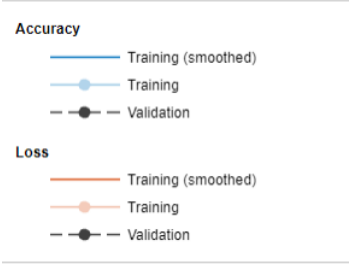
(q)



(r)



(s)



(t)

FIGURE 4.3: The performance of the CNN models visualising training $(accuracy/loss)$ and validation $(accuracy/loss)$ curves. (**a**) AlexNet. (**b**) DarkNet19. (**c**) Darknet53. (**d**) DenseNet201. (**e**) EfficientNetb0. (**f**) Inceptionv1. (**g**) Inceptionv3. (**h**) InceptionResv2. (**i**) MobileNetv2. (**j**) NasnetLarge. (**k**) NasnetMobile. (**l**) ResNet18. (**m**) ResNet50. (**n**) ResNet101. (**o**) ShuffleNet. (**p**) SqueezeNet. (**q**) Vgg16. (**r**) Vgg19. (**s**) Xception. (**t**) Legends.

TABLE 4.1: Characteristics of the pre-trained CNN architectures adopted in the study. $*$: the NasnetLarge and NasnetLarge networks do not contain of a linear sequence of modules.

Network	**#Layers**	**#Learnable Layers**	**Network Size (MB)**	**Input Image Size**	**#Para.(M)**
AlexNet [48]	25	8	227	227×227	61
Vgg16 [68]	41	16	515	224×224	138
Vgg19 [68]	47	19	535	224×224	144
GoogleNet (Inceptionv1) [67]	144	22	27	224×224	7
Inceptionv3 [70]	315	48	89	299×299	23.9
ResNet18 [72]	71	18	44	224×224	11.7
ResNet50 [72]	177	50	96	224×224	25.6
ResNet101 [72]	347	101	167	224×224	44.6
InceptionResv2 [69]	824	164	209	299×299	55.9
Xception [74]	170	71	85	299×299	22.9
DenseNet201 [71]	708	201	77	224×224	20
MobileNetv2 [73]	154	53	13	224×224	3.5
ShuffleNet [75]	172	50	5.4	224×224	1.4
NasnetMobile [76]	913	*	20	224×224	5.3
NasnetLarge [76]	1243	*	332	331×331	88.9
DarkNet19 [77]	64	19	78	256×256	20.8
DarkNet53 [79]	184	53	155	256×256	41.6
EfficientNetB0 [78]	290	82	20	224×224	5.3
SqueezeNet [80]	68	18	5.2	227×227	1.24

Results of the Experiments and Discussion

To examine the classification performance of the models, nine evaluation metrics widely used in classification tasks are used, including, accuracy, classification error, precision, sensitivity, specificity, f1-score, false-positive rate, false-negative rate, and Matthews correlation coefficient. Table 4.2 depicts the evaluation performance of the nineteen CNN models describing the average value and the standard deviation of a specific criterion over the five folds. This study reveals the high evaluation performance of the CNN models for melanoma diagnosis employing a balanced number of dermoscopic images through a thorough analysis of nineteen pre-trained CNNs using specific parameter configuration and learning techniques for the networks.

As shown in Table 4.2, the ResNet101 model reported the best average test accuracy and MCC with 94.34% and 88.96%, respectively, compared to other CNN models. The highest F1-score with a value of 93.96% has been attained by Densenet201, followed by ResNet101 with a value of 93.89%. Furthermore, Inceptionv3 achieved the highest specificity and precision values with 96.8% and 96.11%, followed by 96% specificity achieved by MobileNetv2 and 95.36% precision achieved by ResNet101. DenseNet201 produces the highest sensitivity of 93.47%, followed by 92.86% reported in ResNet101. It can also be noticed that Inceptionv3 attained the lowest FPR of 3.2%, while DenseNet201 revealed the lowest FNR of 6.53%, and the smallest error, 5.66%, is reported by ResNet101. According to the minimum number of parameters, SqueezeNet has 1.24 million parameters which is the optimal number compared to other CNN models. Table 4.2 also explores the deviation among the accuracy reported from the five folds, exposing the difficulty of recognising the best model based on the variation of the accuracy in the five folds. Likely, Table 4.3 and Figure 4.4 show that there is no superior CNN model over others due to the lack of a CNN model that achieves the best accuracy through the five folds. This would lead to difficulty in selecting the best model while considering other criteria.

Figure 4.5 exhibits the trade-off and conflict among the evaluation criteria of the nineteen CNN models. For instance, a trade-off between sensitivity (true positive rate) and specificity (true negative rate) should be considered, where DenseNet201 reports the highest sensitivity, whereas Inceptionv3 attains the highest specificity. Precision is also independent and has a trade-off with accuracy. Accuracy is the degree of veracity, while precision is the degree of reproducibility. That means it is possible to be very precise but not very accurate, and it is also possible to be accurate without being precise. The best quality detection is both accurate and precise. Inceptionv3 achieves the highest precision, whereas Resnet101 reveals the best accuracy. It should also produce a trade-off between FNR and FPR, where Inceptionv3 reported the lowest FPR, while DenseNet201 reported the lowest FNR. Thus, it is crucial to make a trade-off between the models that could achieve the optimal diagnosis by reducing the number of negative

cases falsely diagnosed positive and the models that could reach the optimal diagnosis by reducing the number of positive instances falsely diagnosed negative. F1-Score is also needed to achieve a balance between precision and sensitivity, where Densenet201 provides the best F1-Score, followed by Resnet101. For the number of parameters required to determine the network complexity, SqueezeNet has the lighter network architecture compared to VGG19, which has the largest network architecture. Although SqueezeNet is optimal in terms of network complexity, it still shows a moderate-low accuracy performance through the five folds shown in Figure 4.4. Also, there is a conflict between the criteria that are required to be minimised (such as FNR, FPR, Err, and the number of parameters) and the criteria targeted to be maximised (such as Acc, Sen, Spe, Pre, F1-Score, Mathew).

From Figure 4.5, it can also be noticed that there is no superior CNN model due to the conflict among evaluation criteria and the difficulty to optimise all criteria simultaneously. Hence, selecting the best deep learning model for automated melanoma diagnosis considering multiple conflicted criteria is a difficult task due to the variance of the criteria's significance, the conflict among these criteria, and the trade-off among them. Therefore, benchmarking CNN architectures for melanoma detection is crucial for selecting the optimal model, and achieving the trade-off among the ten pre-defined evaluation criteria. Multiple Criteria Decision-Making method (MCDM) [205, 206] is targeted to apply and rank the nineteen models according to their performance considering the trade-off among the criteria. Thus, the best-selected networks could be easily adopted to construct an ensemble learning system for melanoma diagnosis or even use the optimal network to construct a system using a single model.

TABLE 4.2: The evaluation performance of the nineteen CNN models describing the mean value (m) ± standard deviation (s) of a specific criterion over the five folds.

Network	**mACC± sACC**	**mSen± sSen**	**mSpe± sSpe**	**mF1± sF1**	**mFNR± sFNR**	**mFPR± sFPR**	**mPre± sPre**	**mMathew± sMathew**	**mErr± sErr**
AlexNet [48]	87.07±5.11	84.9±10.95	89.2±3.7	86.39±6.28	15.1±10.95	10.8±3.7	88.57±3.49	74.6±9.87	12.93±5.11
Vgg16 [68]	89.7±6.23	86.94±9.34	92.4±6.5	89.18±6.9	13.06±9.34	7.6±6.5	91.98±6.4	79.76±12.1	10.3±6.23
Vgg19 [68]	87.37±7.01	83.27±11	91.4±10.33	86.58±7.76	16.73±11	8.6±10.33	91.29±9.02	75.64±13.38	12.63±7.01
GoogleNet (Inceptionv1) [67]	87.78±5.87	87.55±8.88	88±11	87.65±5.92	12.45±8.88	12±11	88.71±9.11	76.3±11.22	12.22±5.87
Inceptionv3 [70]	92.93±8.01	88.98±11.82	96.8±4.32	92.29±9.05	11.02±11.82	3.2±4.32	96.11±5.49	86.18±15.55	7.07±8.01
ResNet18 [72]	90±5.68	89.18±4.71	90.8±10.13	89.97±5.38	10.82±4.71	9.2±10.13	91.23±9.32	80.41±11.34	10±5.68
ResNet50 [72]	92.42±7.07	90.2±11.24	94.6±5.22	91.95±7.81	9.8±11.24	5.4±5.22	94.2±5.69	85.21±13.85	7.58±7.07
ResNet101 [72]	94.34±7.28	92.86±12.14	95.8±3.19	93.89±8.26	7.14±12.14	4.2±3.19	95.36±3.94	88.96±14.02	5.66±7.28
InceptionResv2 [69]	90.3±7.96	88.57±10.54	92±5.79	89.87±8.63	11.43±10.54	8±5.79	91.34±6.77	80.71±15.82	9.7±7.96
Xception [74]	88.99±6.79	90±7.85	88±8.8	89.02±6.68	10±7.85	12±8.8	88.39±8.06	78.3±13.59	11.01±6.79
DenseNet201 [71]	93.94±4.97	93.47±3.86	94.4±8.73	93.96±4.7	6.53±3.86	5.6±8.73	94.75±7.64	88.15±9.6	6.06±4.97
MobileNetv2 [73]	90.81±7.24	85.51±11.95	96±3.39	89.9±8.14	14.49±11.95	4±3.39	95.23±4.32	82.25±13.98	9.19±7.24
ShuffleNet [75]	86.06±6.84	80.61±9.16	91.4±14.33	85.24±6.46	19.39±9.16	8.6±14.33	91.99±11.52	73.6±13.19	13.94±6.84
NasnetMobile [76]	86.57±6.47	80.82±12.71	92.2±5.97	85.25±7.87	19.18±12.71	7.8±5.97	91.28±5.36	74.09±12.12	13.43±6.47
NasnetLarge [76]	91.31±7.08	88.16±7.24	94.4±7.7	90.96±7.22	11.84±7.24	5.6±7.7	94.04±7.9	82.84±14.17	8.69±7.08
DarkNet19 [77]	86.77±4.14	81.02±5.43	92.4±3.36	85.79±4.65	18.98±5.43	7.6±3.36	91.22±3.95	73.98±8.15	13.23±4.14
DarkNet53 [79]	89.19±6.15	83.88±9.88	94.4±2.97	88.26±7.22	16.12±9.88	5.6±2.97	93.42±4	78.87±11.79	10.81±6.15
EfficientNetB0 [78]	86.87±3.44	85.31±3.86	88.4±4.88	86.56±3.51	14.69±3.86	11.6±4.88	87.96±4.91	73.86±7.03	13.13±3.44
SqueezeNet [80]	84.65±2.38	86.73±4.95	82.6±6.19	84.83±2.31	13.27±4.95	17.4±6.19	83.34±4.8	69.66±4.75	15.35±2.38

TABLE 4.3: The obtained accuracies over five folds in the nineteen CNN models.

Model	Fold1	Fold2	Fold3	Fold4	Fold5
AlexNet	78.28	89.9	86.87	90.4	89.9
Vgg16	82.32	86.87	86.87	95.96	96.46
Vgg19	79.8	80.81	90.4	89.39	96.46
Inceptionv1	82.32	84.85	83.84	91.92	95.96
Inceptionv3	79.8	90.91	96.97	99.49	97.47
ResNet18	82.83	85.35	92.93	92.42	96.46
ResNet50	81.31	90.91	92.93	97.98	98.99
ResNet101	81.82	94.44	96.97	98.99	99.49
InceptionResv2	77.27	88.89	93.43	93.94	97.98
Xception	78.28	86.87	90.91	93.43	95.45
DenseNet201	86.36	91.94	96.46	97.98	97.47
MobileNetv2	81.31	86.87	89.9	97.47	98.48
ShuffleNet	77.27	83.84	84.85	88.38	95.96
NasnetMobile	78.28	86.36	85.86	85.86	96.46
NasnetLarge	80.3	88.89	92.93	96.46	97.98
DarkNet19	81.31	85.86	84.85	90.91	90.91
DarkNet53	79.29	87.37	91.41	93.94	93.94
EfficientNetB0	84.34	83.84	85.35	88.89	91.92
SqueezeNet	82.32	84.85	82.32	85.86	87.88

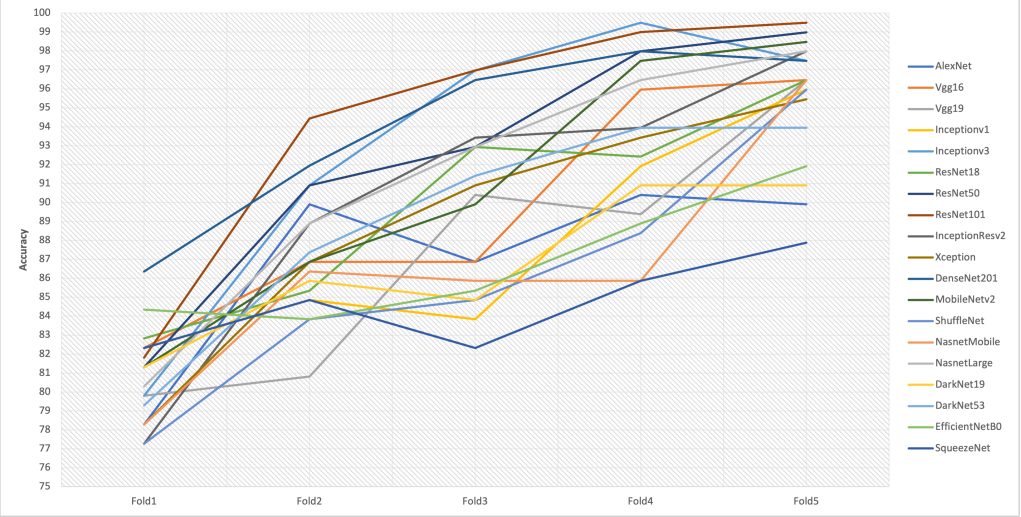


FIGURE 4.4: The obtained accuracies over five folds in the nineteen CNN models. It shows that there is no superior CNN model over others due to the lack of a CNN model that achieves the best accuracies through the five folds. This would lead to difficulty selecting the best model while considering another conflicting criterion like the network complexity.

FIGURE 4.5: The mean value over the five folds for specific evaluation criteria, along with the number of parameters (the network complexity). No single model achieves the best performance in all evaluation criteria. If a CNN model achieves the best evaluation performance in some evaluation criteria, it may fail to gain superior performance in the remaining criteria.

To achieve the goal of the presented study by generating a ranking list for CNN models and selecting the optimal solution, the PROMETHEE method [205] is applied considering the nineteen alternatives (CNN models) and ten criteria. To further validate the decision made by PROMETHEE, the VIKOR approach [206] is also applied using the same data setting and configuration. First, the decision matrix DM is constructed using m alternatives, in this case study is 19, and k criteria, in this case study is 10 producing DM of size 19×10. The criteria are then classified into two categories according to the required optimisation strategy. The first category includes the criteria that require minimisation, including classification error, false-positive rate, false-negative rate and the number of parameters, known as non-beneficial criteria. Unlikely, the second category includes the criteria that require maximisation, including accuracy, sensitivity, specificity, precision, F1-score and MCC, known as beneficial criteria. The equations 4.24 and 4.25 defined below are used to normalising the non-beneficial and beneficial criteria, respectively. The normalised criteria are shown in Table 4.4.

$$\bar{x}_{ij} = \frac{x_j^{\min}}{x_{ij}} \tag{4.24}$$

$$\bar{x}_{ij} = \frac{x_{ij}}{x_j^{\max}} \tag{4.25}$$

x_{ij} refers to the entries of the decision matrix DM, where $i = 1, \ldots, m$, $j = 1, \ldots, k$, k represents the number of alternatives (nineteen CNN models), and m defines the number of criteria (ten criteria).

To measure the weights of criteria, the entropy method [207] is exploited and applied on the normalised DM producing the weight values of 0.964825438, 0.804398756, 0.985470611, 0.951881312, -1.420375792, -1.473036988, 1.02152041, 0.49110277, -1.294287661, -0.031498856 for accuracy, sensitivity, specificity, F1-score, FNR, FPR, precision, MCC, classification error and the number of parameters, respectively. The obtained weights, along with the normalised DM, are used to make the optimal selection using the PROMETHEE method [205]. The equations used to measure the ranking list are described earlier in Section 4.2.2. The threshold function has been used as a preference function (0 if d $\leq$ 0 and 1 if d $\geq$ 0), which is required in Step **2** in the stepwise procedure of PROMETHEE. To calculate the complete ranking list, $\phi(a)$ represents the outranking flow for each alternative, as shown in Table 4.5. The highest $\phi(a)$ value indicates the compromised solution, which could be chosen as the optimal model. PROMETHEE reports a value of 150.84, the highest $\phi(a)$ for the ResNet101 CNN model and 133.24 as the second-best value for the DenseNet201 model.

To validate the model selection made by PROMETHEE, VIKOR [206] method is also applied considering the same weights and the same DM. Unlike PROMETHEE, the lowest Q value in VIKOR indicates the compromised solution, which could be chosen as the optimal model, shown in Table 4.5. VIKOR reports a value of 0, the lowest Q for the ResNet101 CNN model and 0.079 as the second-lowest value for the DenseNet201 model. Thus, the mathematical consistency of the judgements coming out of PROMETHEE II has been tested and proven. Hence, the effectiveness of the

model ranking produced by PROMETHEE II has been validated by demonstrating the agreement between two different statistical methods, considering the same conflicting criteria.

To provide a direct and explicit comparison between the two decision-making methods, PROMETHEE and VIKOR, Table 4.6 elaborates the optimal CNN model selection in both approaches. It can be noticed that until the seventh rank, the two methods have a similar decision for the optimal CNN model selection. Likewise, the ranks 10, 11, 12, 13, 15, 18 and 19 provide the exact model recommendation by both approaches. On the other hand, the decision made by methods has slightly different priorities in the 8, 9, 14, 16 and 17 levels. The suggested framework's findings show that the best model selection decision based on numerous conflict factors is robust and reliable.

This work developed a new multi-criteria decision-making methodology that aids in assessing the criteria that influence the decision to choose a specific CNN model, prioritising the models, and selecting the best model. When software developers need to find an effective CNN model that meets specified requirements for constructing a robust CAD system, the proposed approach of revealing the CNN models' priorities would be beneficial and valuable. Finally, the case study presented here may provide and draw a new line in the evaluation and benchmark of the deep learning models for various diseases. Although the proposed benchmarking framework has made progress in benchmarking the models used for melanoma diagnosis from dermoscopy images, there is still space for improvement in research work. In future work, studying the effect of model selection considering different criteria is aimed. The criteria that are to be considered include i) training the models under several transfer learning scenarios and data augmentation strategies, ii) exploring the impact of several optimisation schemes, and iii) testing various class balancing and weighting techniques. Training the models on several datasets targeting the effect of variation among datasets could also be considered.

TABLE 4.4: Normalised decision matrix. Alter:: refers to alternative and Cr. refers to criterion.

Alter./ Cr.	**ACC**	**Sen**	**Spe**	**F1-Score**	**FNR**	**FPR**	**Pre**	**MCC**	**Err**	**Para**
AlexNet	0.9229	0.9083	0.9215	0.9194	0.4325	0.2963	0.9215	0.8386	0.4377	0.0203
Vgg16	0.9508	0.9301	0.9545	0.9491	0.5000	0.4211	0.9570	0.8966	0.5495	0.0090
Vgg19	0.9261	0.8909	0.9442	0.9215	0.3903	0.3721	0.9498	0.8503	0.4481	0.0086
Inceptionv1	0.9305	0.9367	0.9091	0.9328	0.5245	0.2667	0.9230	0.8577	0.4632	0.1771
Inceptionv3	0.9851	0.9520	1.0000	0.9822	0.5926	1.0000	1.0000	0.9688	0.8006	0.0519
ResNet18	0.9540	0.9541	0.9380	0.9575	0.6035	0.3478	0.9492	0.9039	0.5660	0.1060
ResNet50	0.9796	0.9650	0.9773	0.9786	0.6663	0.5926	0.9801	0.9578	0.7467	0.0484
ResNet101	1.0000	0.9935	0.9897	0.9993	0.9146	0.7619	0.9922	1.0000	1.0000	0.0278
InceptionResv2	0.9572	0.9476	0.9504	0.9565	0.5713	0.4000	0.9504	0.9073	0.5835	0.0222
Xception	0.9433	0.9629	0.9091	0.9474	0.6530	0.2667	0.9197	0.8802	0.5141	0.0541
DenseNet201	0.9958	1.0000	0.9752	1.0000	1.0000	0.5714	0.9858	0.9909	0.9340	0.0620
MobileNetv2	0.9626	0.9148	0.9917	0.9568	0.4507	0.8000	0.9908	0.9246	0.6159	0.3543
ShuffleNet	0.9122	0.8624	0.9442	0.9072	0.3368	0.3721	0.9571	0.8273	0.4060	0.8857
NasnetMobile	0.9176	0.8647	0.9525	0.9073	0.3405	0.4103	0.9497	0.8328	0.4214	0.2340
NasnetLarge	0.9679	0.9432	0.9752	0.9681	0.5515	0.5714	0.9785	0.9312	0.6513	0.0139
DarkNet19	0.9198	0.8668	0.9545	0.9130	0.3440	0.4211	0.9491	0.8316	0.4278	0.0596
DarkNet53	0.9454	0.8974	0.9752	0.9393	0.4051	0.5714	0.9720	0.8866	0.5236	0.0298
EfficientNetB0	0.9208	0.9127	0.9132	0.9212	0.4445	0.2759	0.9152	0.8303	0.4311	0.2340
SqueezeNet	0.8973	0.9279	0.8533	0.9028	0.4921	0.1839	0.8671	0.7830	0.3687	1.0000

TABLE 4.5: Ranking for decision making represented by the values of ϕ in PROMETHEE and Q in VIKOR. The highest ϕ value is the best, whereas the lowest Q is the best.

Model	ϕ**: PROMETHEE**	Q**: VIKOR**	**PROMETHEE**	**VIKOR**
AlexNet	-86.54004365	0.78423285	15	15
Vgg16	16.31877628	0.51048488	8	9
Vgg19	-63.8124359	0.74766659	13	13
Inceptionv1	-57.19966687	0.68096691	12	12
Inceptionv3	132.2050634	0.18466346	3	3
ResNet18	15.25546934	0.4614654	9	8
ResNet50	115.1633097	0.21251132	4	4
ResNet101	150.8418215	0	1	1
InceptionResv2	28.425464	0.4496109	7	7
Xception	-29.98203689	0.60787425	11	11
DenseNet201	133.2355605	0.07998389	2	2
MobileNetv2	72.89230795	0.42167181	6	6
ShuffleNet	-106.9819714	0.8594925	18	18
NasnetMobile	-89.20093646	0.84854	16	17
NasnetLarge	73.3193101	0.33461685	5	5
DarkNet19	-76.30565263	0.81073772	14	16
DarkNet53	1.456682009	0.56957337	10	10
EfficientNetB0	-95.9301979	0.78239429	17	14
SqueezeNet	-133.1608231	1	19	19

TABLE 4.6: Optimal CNN model selection in PROPMETHEE versus VIKOR approach.

Model Rank	**PROPMETHEE**	**VIKOR**
1	ResNet101	ResNet101
2	DenseNet201	DenseNet201
3	Inceptionv3	Inceptionv3
4	ResNet50	ResNet50
5	NasnetLarge	NasnetLarge
6	MobileNetv2	MobileNetv2
7	InceptionResv2	InceptionResv2
8	Vgg16	ResNet18
9	ResNet18	Vgg16
10	DarkNet53	DarkNet53
11	Xception	Xception
12	Inceptionv1	Inceptionv1
13	Vgg19	Vgg19
14	DarkNet19	EfficientNetB0
15	AlexNet	AlexNet
16	NasnetMobile	DarkNet19
17	EfficientNetB0	NasnetMobile
18	ShuffleNet	ShuffleNet
19	SqueezeNet	SqueezeNet

4.4 Conclusions

Medical diagnostics tools based on deep learning of medical images are becoming more widely recognised as clinically relevant AI-based solutions. However, developing appropriate deep neural network models and training strategies for clinical uses is a research area that needs to be investigated. The inaccurate selection of melanoma diagnosis models could be costly to medical organisations, especially when more accurate and efficient diagnosis models are urgently needed. This study investigated the performance of some of these networks for melanoma diagnosis utilising dermoscopic images after a thorough evaluation of nineteen pre-trained CNNs using particular evaluation criteria, parameter settings and training strategies. An MCDM-based methodology is presented for evaluating, benchmarking, and ranking melanoma diagnostic models and selecting the most optimal model. The study findings would help in the model selection, designing quick and reliable diagnostic tools based on image data, and contributing to the development of more accurate and efficient point-of-care diagnostic systems.

Chapter 5

Classification of Skin Lesions Related to Melanoma From Dermoscopic and Clinical Images via Seven-point Checklist Criteria

Following the models reported in the previous chapter, which are developed to detect melanoma-related lesions from only dermoscopic images, this chapter studies melanoma detection using a classification-based deep learning technique via seven-point checklist criteria from both clinical and dermoscopic images. The deliverables of the research work reported in this chapter were a full paper published at European Workshop on Visual Information Processing conference (EUVIP) in Alzahrani et al. [194], and a two-page short paper presented as a poster at International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI) in Alzahrani et al. [209]. The contribution of the first author to work presented in this chapter is the conceptualisation of the idea, proposing the methodology, development of the model, data and results analysis and writing up. The labelled dataset used for training and testing was provided by the authors of [210].

5.1 Introduction

In clinical practice, dermatologists typically assess skin lesions using the seven-point checklist method [10] or the ABCDE rule [110]. These methods are considered the most commonly recommended and accepted skin assessment strategies [36]. In the seven-point checklist rule, melanoma suspicion is greater for lesions scoring 3 points or exceeding 3, while low-suspicion lesions should be carefully screened and monitored for changes for eight weeks. The seven-point checklist was established by Argenziano et al. [10] for the dermoscopic differentiation between benign and malignant lesions. The definition of the seven-point checklist criteria can be briefly described as follows:

1. **Atypical Pigment Network (PN):** Reticular lines, heterogeneous for colour and thickness, asymmetrically distributed within the lesion.

2. **Blue Whitish Veil (BWV):** Structureless blue blotches with an overlying whitish haze.

3. **Vascular Structure (VS):** Linear, dotted globular vessels (polymorphic vessels), irregularly distributed.

4. **Irregular Pigmentation (PIG):** Structureless area different in size and colour (black, brown or gray) irregularly distributed.

5. **Irregular Streaks (STR):** Radial streaks and pseudopods located at the lesion edge due to the melanoma radial growth phase.

6. **Irregular Dots and Globules (Dag):** Dots (less than 0.1 mm) and globules (larger than 0.1 mm), irregular in colour, size, shape and distribution.

7. **Regression Structures (RS):** White scar-like depigmentation or peppering (multiple scattered blue-gray granules within a hypo-pigmented background).

The principle of this method [10] establishes three major dermoscopic criteria (2 points each) and 4 minor criteria (1 point each) for lesions assessment as shown in

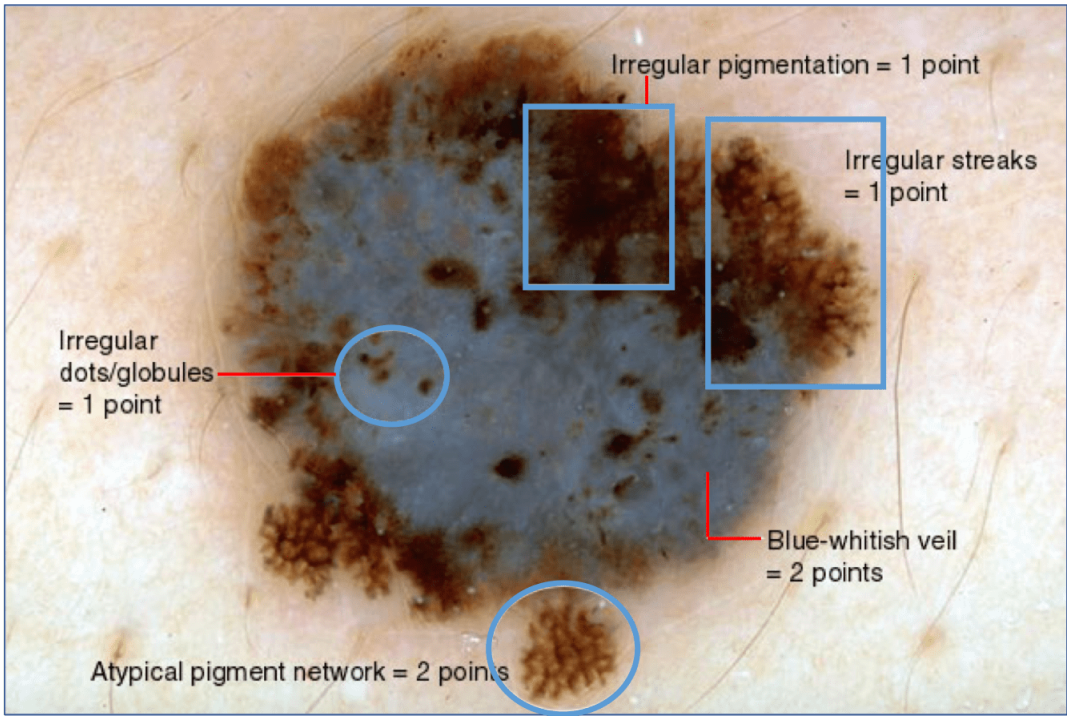
Table 5.1[10]. Image examples which are diagnosed using seven-point checklist criteria can be shown in Figure 5.1.

TABLE 5.1: Seven-point checklist criteria

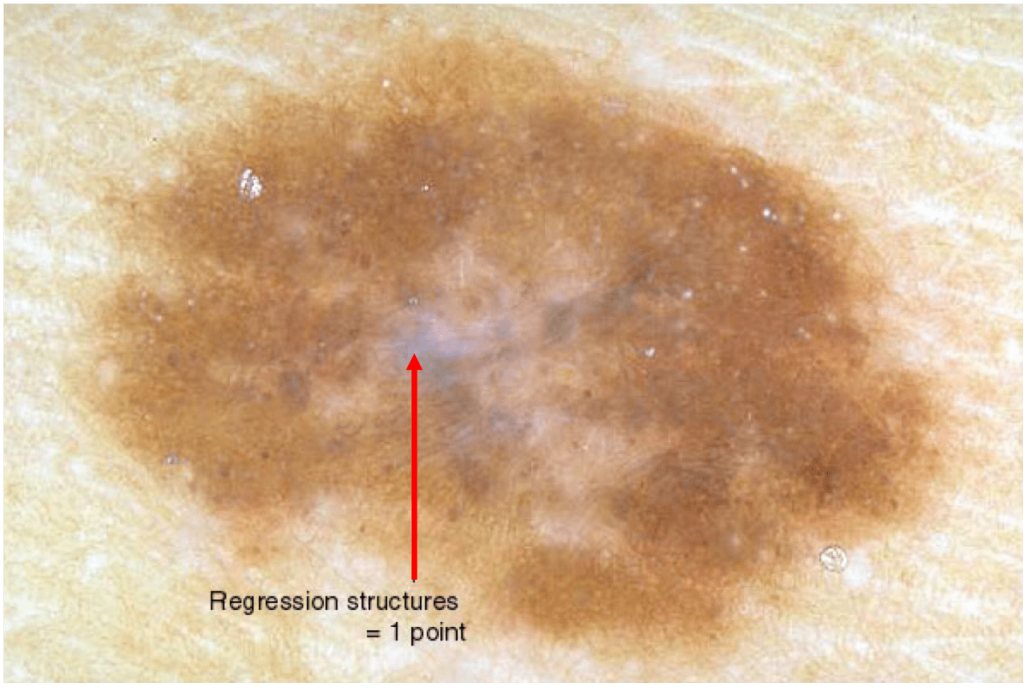
Criteria	Classes	7-Point Score
Major Criteria		
Pigment Network (PN)	absent (0), typical (0), atypical (2)	2
Blue Whitish Veil (BWV)	absent (0), present (2)	2
Vascular Structure (VS)	absent (0), regular (0), irregular (2)	2
Minor Criteria		
Pigmentation (PIG)	absent (0), regular (0), irregular (1)	1
Streaks (STR)	absent (0), regular (0), irregular (1)	1
Dots and Globules (Dag)	absent (0), regular (0), irregular (1)	1
Regression Structures (RS)	absent (0), present (1)	1

5.2 Related Work

Computer-Aided Diagnosis (CAD) systems are introduced in clinical practice for dermoscopic images to provide an automatic and quantitative assessment of the skin lesion to help clinicians with diagnosis. Many automated skin lesion detection methods in the literature considered detecting a single criterion from the 7-point checklist. Mirzaalian et al. [211] presented a method to detect only streaks by enhancing streaks using Hessian-based tubular filters from 99 images. Moreover, Madooie et al. [212] proposed a system to detect the presence of blue-white veils by converting image regions to a discrete set of Munsell colours. Furthermore, Fabbrocini et al. [213] developed a machine learning-based system to detect all 7-point checklist criteria by implementing a separate pipeline which considers each criterion's characteristics separately. However, each pipeline introduces complexity and requires tuning of hyperparameters. In addition, Wadhawan et al. [214] proposed a machine learning-based method to detect all 7- point checklist criteria by engineering image features manually, which is a time-consuming and complicated task.



(a)



(b)

FIGURE 5.1: (a) Image example is diagnosed with Melanoma given 7-points score = 7 (b) Image example is diagnosed with Non-Melanoma given 7-points score = 1. Adapted from [10]

Many deep learning (pattern analysis) methods have been reported in the literature to analyse skin lesion images. The pattern analysis methods do not consider the seven-point checklist criteria or ABCDE rule; instead, these methods work on extracting the features of lesions automatically and use the extracted features to infer melanoma diagnosis. Codella et al. presented a hybrid method, integrating sparse coding, convolutional neural network (CNN) and support vector machines (SVMs) to detect melanoma [54]. Recently, Codella et al. established a method combining recent developments in deep learning and machine learning approaches for skin lesion segmentation and classification [215]. Kawahara et al. used a fully convolutional network to extract multi-scale features for melanoma detection and recognition [216]. Yu et al. applied a very deep residual network to distinguish melanoma from non-melanoma lesions [217]. Esteva et al. [125] adopted a pre-trained GoogleNet Inception v3 CNN model to classify skin cancers. Haenssle et al. [127] utilised a deep convolutional neural network to classify a binary diagnostic category of dermoscopy images of melanocytic images. Dorj et al. [218] developed SVM with a deep convolutional neural network approach for the classification of four diagnostic categories of clinical skin cancer images. Han et al. [140] used a deep convolutional neural network to classify the clinical images of 12 skin diseases.

From literature, it has been found by experienced dermatologists that the 7-point checklist gives higher sensitivity but lower specificity than some pattern analysis methods [219, 220]. This indicates the limitations in both approaches and reports a trade-off between the two assessment strategies, thereby motivating more investigation and analysis of both approaches. Furthermore, although the pattern analysis and seven-point checklist criteria diagnostic procedures are different, the seven criteria are basically based on the criteria exploited in the process of pattern analysis and interpretations [221]. In addition, detecting and localising these criteria can help with more understandable and interpretable diagnostic procedures, such as recognising the presence of malignant features and retrieving images that satisfy specific criteria.

In this research work, a deep learning-based method has been proposed to predict the 7-points checklist criteria and diagnose melanoma where the lesion features are designed automatically. Multiple input convolutional neural networks (CNNs) considering clinical and dermoscopic images as inputs have been developed. Incorporating 7-point checklist criteria with CNN and learning the proposed model using difficult and non-standardised images (clinical images) may aid with leveraging the reliability of melanoma diagnosis.

5.3 Materials and Methodology

5.3.1 Materials

A publicly available dataset [210] for evaluating computerised image-based prediction of the 7-point skin lesion malignancy checklist has been used in the experiments. The dataset includes over 2000 clinical and dermoscopy colour images for training and evaluating computer-aided diagnosis (CAD) systems. Dermoscopic images are taken with a dermatoscope and offer a standardised field of view. The dermoscopic images are captured under controlled conditions like standard illumination, lighting and contrast. Clinical image acquisition is carried out under less standardised conditions, such as various fields of view and containing image artefacts.

5.3.2 Methodology

Convolutional neural networks (CNNs) can be adopted as feature learning algorithms because the convolutional neural layers have a credible ability to detect good features in the images and form hierarchies of nonlinear features where their complexity grows while going deeper through the network. The main idea of CNN is stacking such deep hierarchies of nonlinear features. For images, it can be mathematically shown that edges and blobs are the best features that can be extracted in the earlier layers. To generate features containing more information, earlier features (edges and blobs) are transformed

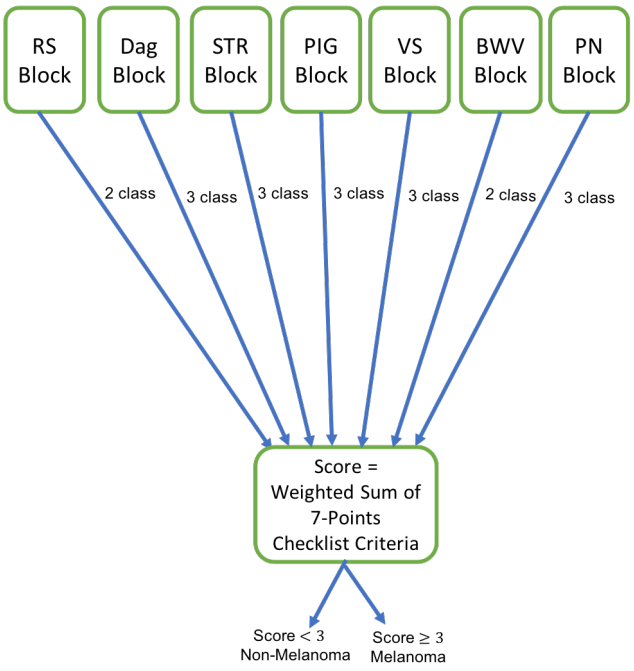
again to obtain more complex features that consist of more information to discriminate among the classes.

The Block diagram of the proposed method is shown in Figure 5.2. The upper subfigure shows the abstract level of the developed system, while the lower subfigure shows the details of each classification block. The diagram has seven classification blocks, one for each lesion attribute, to predict the attribute label. The predicated attribute label is taken into consideration along with the other labels, and their weights are summed up to produce the diagnosis score according to 7- point checklist criteria. If the obtained score is equal to or greater than three, then the decision taken by the network results in class 1, indicating to melanoma case; otherwise, results in class 0 indicating non-melanoma.

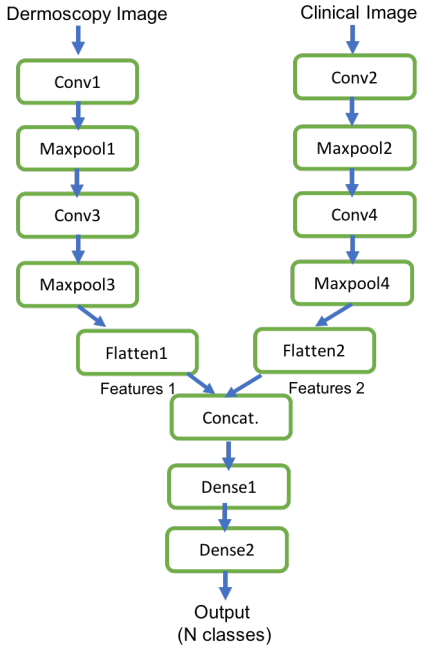
The architecture of the implemented convolutional network of each classification block is described in Table 5.2. The last column shows the filters' size and the max-pooling window size in each layer. Both networks have the same architecture consisting of two convolutional layers, each followed by a max-pooling layer. The extracted features from two networks are concatenated together and passed into dense layers (fully connected layers) in order to identify the lesion attribute from the input image. The deep network was trained through 300 epochs with a cross-entropy loss function. The training of the model is set to 300 epochs where the validation error stops improving after this epoch, and the best weighting values are maintained. Stochastic gradient descent (SGD) with a momentum optimisation algorithm having a learning rate of 0.001 and a momentum parameter of 0.9 is used to train the networks by updating the weights.

5.4 Results and Discussions

The data are randomly divided into 60% training and the remaining for testing. All images were resized to $256 \times 256 \times 3$. The performance of the proposed method for lesion identification and melanoma diagnosis, when compared with the ground truth, was evaluated using many evaluation metrics, including accuracy, sensitivity, and specificity.



(a)



(b)

FIGURE 5.2: (a) Block diagram of proposed skin lesion attribute classification. PN: Pigment Network, BWV: Blue Whitish Veil, VS: Vascular Structure, PIG: Pigmentation, STR: Streaks, Dag: Dots and Globules, RS: Regression Structures. (b) Details of each skin lesion classification block shown in (a).

TABLE 5.2: Convolutional neural network architecture and its parameters

Layer	Output Shape	Connection	Size
Dermoscopy input1 image	(256, 256, 3)	-	-
Clinical input2 image	(256, 256, 3)	-	-
Conv1	(253, 253, 32)	Dermoscopy input1 image	filter size = (4,4)
Conv2	(253, 253, 32)	Clinical input image2	filter size = (4,4)
MaxPooling1	(127, 127, 32)	Conv1	Maxpool size = (2,2)
MaxPooling2	(127, 127, 32)	Conv2	Maxpool size = (2,2)
Conv3	(124, 124, 16)	MaxPooling1	filter size = (4,4)
Conv4	(124, 124, 16)	MaxPooling2	filter size = (4,4)
MaxPooling3	(62, 62, 16)	Conv3	Maxpool size = (2,2)
MaxPooling4	(62, 62, 16)	Conv4	Maxpool size = (2,2)
Flatten1	(1,61504)	MaxPooling3	-
Flatten2	(1,61504)	MaxPooling4	-
Concatenation	(1,123008)	Flatten1 and Flatten2	-
Dense1	(1,100)	Concatenation	-
Dense2	(1,100)	Dense1	-
Output	(N)	Dense2	N= 2 or 3

5.4.1 Baseline

The proposed method has been trained and evaluated on the same dataset used in the state-of-the-art method presented by Kawahara et al. [210], where a multi-model deep learning method to predict 7-point checklist criteria using metadata was developed. The baseline results obtained from the proposed system for lesion detection and melanoma diagnosis are presented in Table 5.3.

TABLE 5.3: Baseline results of skin lesion attribute detection and melanoma diagnosis.

Lesion	Accuracy	Sensitivity	Specificity
Pigment Network (PN)	0.6278	0.5957	0.8092
Blue Whitish Veil (BWV)	0.8051	0.8750	0.5067
Vascular Structure (VS)	0.7316	0.5620	0.8156
Pigmentation (PIG)	0.6278	0.5663	0.7848
Streaks (STR)	0.6759	0.5807	0.7808
Dots and Globules (Dag)	0.5038	0.4832	0.7475
Regression Structures (RS)	0.7190	0.8097	0.4717
Lesion detection-average	0.6701	0.6387	0.7023
Melanoma diagnosis	0.6430	0.5537	0.8926

Madooie et al. [212], who proposed a system to detect the presence of blue-white veils (BWV) based on colour analysis, reported a sensitivity of 71%, while the proposed system is able to detect BWV lesions, achieving a sensitivity of 87.5%. Furthermore, Wadhawan et al. [214], who proposed a method based on hand-crafted feature extraction, reported a sensitivity of 79.5% and 64.2% on detecting BWV and RS, respectively. In contrast, the proposed method achieves a sensitivity of 87.5% and 80.79% for identifying the same lesions. Moreover, Kawahara et al. [210] reported results better than the baseline results of the proposed method achieving a sensitivity of 60.4% and specificity of 91% for melanoma diagnosis compared to the proposed system that shows a sensitivity of 55.37% and specificity of 89.26%. However, the baseline results have been achieved without considering label imbalance problems or parameter tuning, which have been carefully addressed in their developed system.

5.4.2 Improved Results

This section presents the extended work of the system developed and presented in [194]. The development and improvements carried out on the proposed lesion detection and melanoma diagnosis system are explored as follows. The proposed lesion detection and melanoma diagnosis comprises seven lesion attributes (L1, L2, . . . , L7). Each lesion is passed into four different models (M1, M2, M3, M4), in addition to Alexnet reported as a baseline, to detect a lesion from the seven-point checklist. Along with the other labels, the predicated lesion attribute label (P1, P2, . . . , P7) is taken into consideration, and their weights are summed to generate the diagnosis score according to the criteria of the 7-point checklist. If the score obtained is equal to or greater than three, the decision taken by the network will result in class 1, indicating the case of melanoma; otherwise, it will result in class 0, indicating non-melanoma. Figure 5.3 shows the abstract level block diagram of the proposed system.

The proposed system has been built using five backbone network models, pre-trained over ImageNet [48] dataset. The five convolutional neural network models are

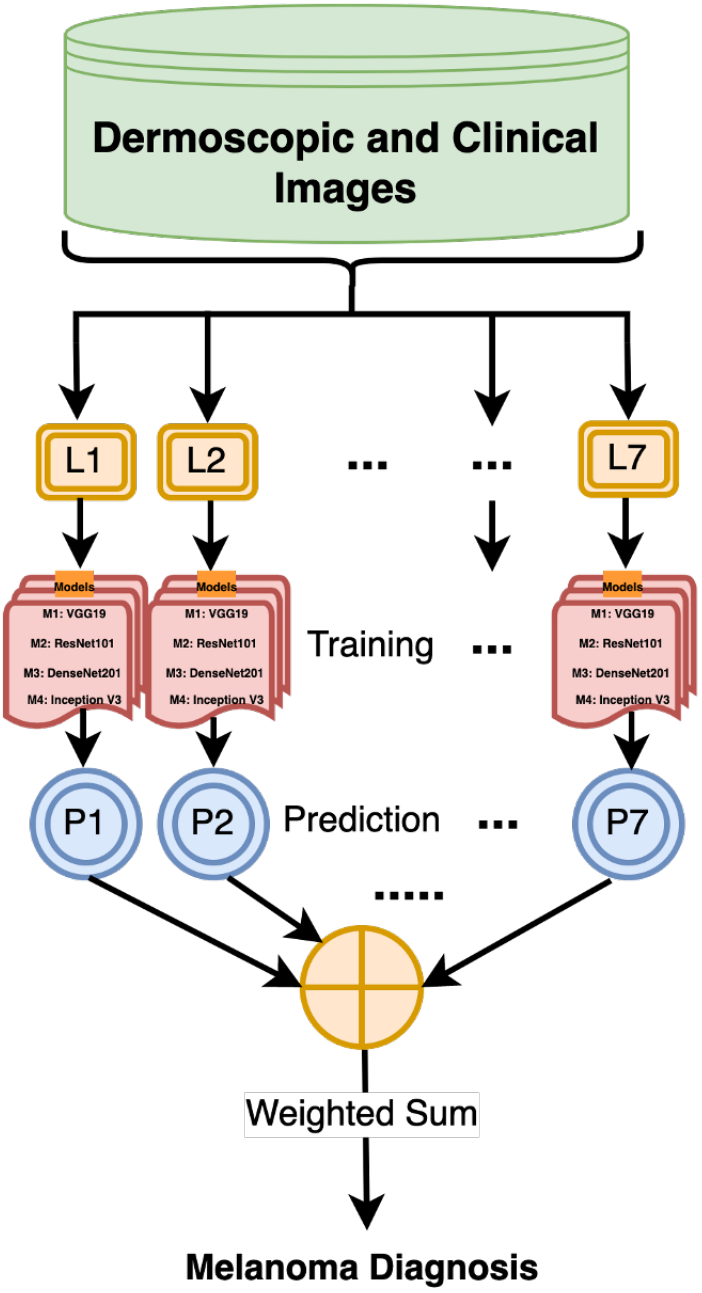


FIGURE 5.3: Block diagram of the proposed melanoma diagnosis system.

AlexNet (as baseline) [48], VGG16 [68], ResNet101 [72], DenseNet201 [71], and Inception V3 [70]. The models are retrained on the dermoscopic and clinical images, where the weights are fine-tuned. Alexnet architecture [48] is adopted as a baseline model. Augmentation is applied to the training images in real-time with horizontal and vertical flipping, rotations, zooms, and random translations. As the data has an imbalanced distribution of positive and negative labels, this issue is addressed by oversampling the minority classes in each batch of training data. The classification is further penalised by imposing an additional cost and weighing up on the minority class during model training. Figures 5.4, 5.5, 5.6, 5.7, and 5.8 depict the performance of the developed system for lesion detection and melanoma diagnosis.

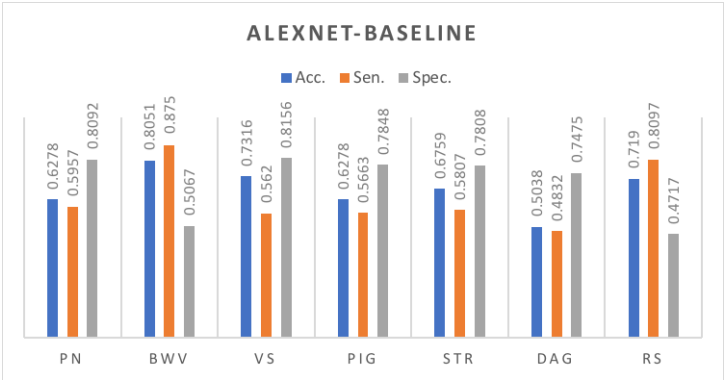


FIGURE 5.4: Seven lesions detection performance in melanoma diagnosis system using Alexnet as the backbone.

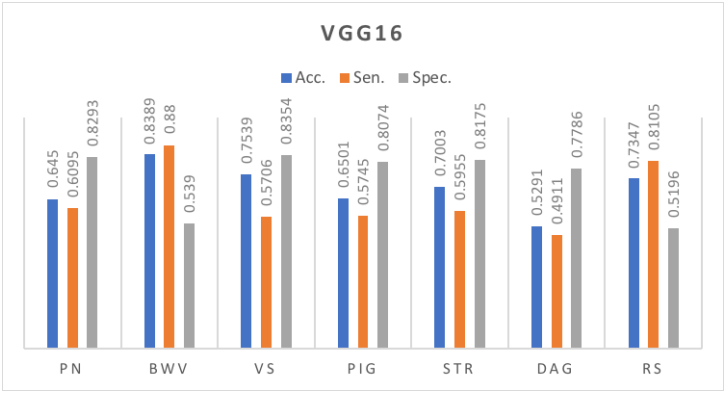


FIGURE 5.5: Seven lesions detection performance in melanoma diagnosis system using VGG16 as the backbone.

Tables 5.4 and 5.5 show the performance of the proposed pipeline system for skin lesion and melanoma detection in terms of accuracy, sensitivity and specificity. Best

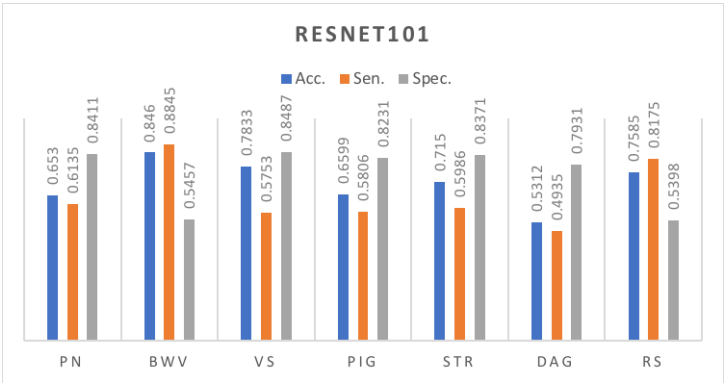


FIGURE 5.6: Seven lesions detection performance in melanoma diagnosis system using Resnet101 as the backbone.

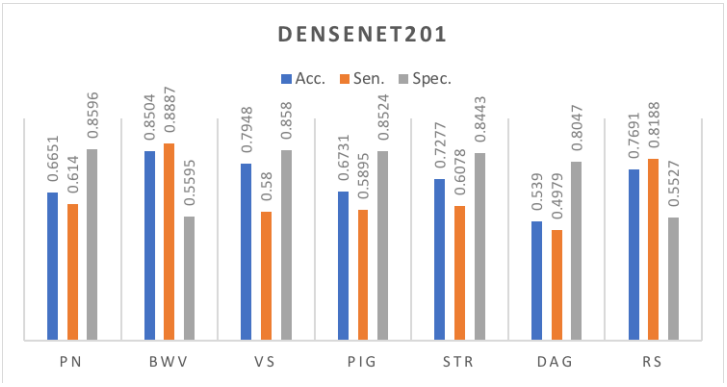


FIGURE 5.7: Seven lesions detection performance in melanoma diagnosis system using Densenet201 as the backbone.

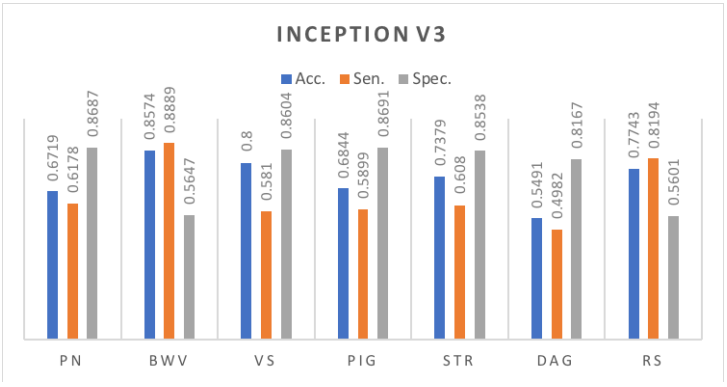


FIGURE 5.8: Seven lesions detection performance in melanoma diagnosis system using Inception V3 as the backbone.

results are obtained from the Inception v3 backbone, producing average accuracy, sensitivity and specificity of 0.7250, 0.6576, and 0.7705; respectively, for lesion detection and 0.7253, 0.6023, and 0.9071; respectively, for melanoma diagnosis.

TABLE 5.4: Average lesion detection performance.

Network	Accuracy	Sensitivity	Specificity
Baseline (AlexNet)	0.6701	0.6387	0.7023
VGG19	0.6931	0.6473	0.7324
ResNet101	0.7067	0.6519	0.7469
DenseNet201	0.717	0.6566	0.7616
Inception V3	**0.725**	**0.6576**	**0.7705**

TABLE 5.5: Average melanoma diagnosis performance.

Network	Accuracy	Sensitivity	Specificity
Baseline (AlexNet)	0.643	0.5537	0.8926
VGG19	0.6722	0.5696	0.8978
ResNet101	0.6989	0.582	0.9003
DenseNet201	0.7121	0.5914	0.9055
Inception V3	**0.7253**	**0.6023**	**0.9071**

Kawahara et al. [210] presented a multi-task system for lesion detection and melanoma diagnosis, achieving an average accuracy of 0.7370, a sensitivity of 0.6620 and specificity of 0.8060 for lesion detection and average accuracy, sensitivity and specificity of 0.7420, 0.6040, 0.9100; respectively for melanoma diagnosis on the same dataset used to evaluate the proposed system. The results of the developed separate pipelines for predicting eight different categories (melanoma diagnosis and seven-point checklist) reveal a close and comparable performance with the multi-task system proposed by the authors of [210]. However, the authors of [210] claimed that the labels of eight categories are not mutually exclusive. Yet, they defined a multi-task loss function with eight terms for eight tasks without considering the dissimilarity among tasks. The dissimilar tasks could affect badly on the generalisation performance of the multi-task system if tested on different unseen datasets. Their method treats all eight tasks with equivalent importance, and therefore it becomes crucial to find a robust strategy to choose the weighting

scheme for each task in the loss function. Moreover, unlike the proposed method here, they considered the lesion detection task as a targeted task instead of an auxiliary task to achieve the final goal, which is melanoma diagnosis.

It is worth mentioning that the studies presented in Chapters 4 and 5 of the analysis aim to examine skin lesions related to melanoma. However, the research questions addressed in these studies differ significantly. Chapter 4 focuses on evaluating and benchmarking nineteen convolutional neural network architectures for melanoma diagnosis through the analysis of dermoscopic images. The goal is to prioritise the convolutional network architectures and select the optimal one based on specific criteria. On the other hand, Chapter 5 aims to determine the potential impact of incorporating seven-point checklist criteria with CNN models and training the models using both clinical and dermoscopic images on the reliability of melanoma diagnosis. Thus, while Chapter 4 endeavours to develop a new benchmarking approach for selecting the optimal architecture for melanoma diagnosis, Chapter 5 aims to develop a new method for the diagnosis of melanoma. It is important to note that the public dataset used in Chapter 4 [126] does not contain the ground truth of the seven-point checklist criteria, making it unsuitable for training and validating the method proposed in Chapter 5. As a result, the results from Chapters 4 and 5 cannot be directly compared due to the differences in research questions and data.

5.5 Conclusions

A new technique for skin lesion detection and melanoma diagnosis from dermoscopy images by combining seven-point checklist criteria with convolutional neural networks has been proposed and implemented successfully. The proposed models have been realised by incorporating automated lesion feature extraction achieved by multi-input CNN considering standardised images (dermoscopy) and non-standardised images (clinical). It is demonstrated that the proposed method performs well in terms of accuracy, sensitivity, and specificity.

Chapter 6

Conclusions and Future work

This chapter presents a summary of the proposed skin lesion analysis methods, the main findings and limitations of the research work, and some potential future research work directions. A summary of the proposed methods is presented in Section 6.1. The main findings and research limitations are reported in Section 6.2. Finally, the further possible research directions are given in Section 6.3.

6.1 Detailed Conclusions

In this thesis, three distinct skin lesion analysis methods based on image classification and regression approaches were proposed and applied to facial, dermoscopic, and clinical images to detect melanoma-related lesions and assess the severity of acne-related lesions. The proposed methods' performance evaluation using several metrics demonstrated that the presented techniques work well in both classification and regression tasks. The utilised public labelled datasets helped to conduct the training and testing of the developed systems, allowing for a fair comparison with the performance of the state-of-the-art methods evaluated on the same public datasets. All of the images used to train the models presented in this thesis were augmented, targeting to artificially increase the samples of image data. The CNNs were adopted to automatically extract and

learn high-level features in the proposed methods, feeding these features to the fully connected layers for a final decision of class prediction.

The first theme of the research presented the acne lesion analysis method using a regression-based deep learning algorithm. Acne vulgaris is the common form of acne that primarily affects adolescents, characterised by an eruption of inflammatory and/or non-inflammatory skin lesions. Accurate evaluation and severity grading of acne plays a significant role in precise treatment for patients. Manual acne examination is typically conducted by dermatologists through visual inspection of the patient skin and counting the number of acne lesions. However, this task costs time and requires excessive effort by dermatologists. Thus, this work presented automated acne counting and severity grading method from facial images. To this end, a multi-scale dilated fully convolutional regressor integrated with an attention mechanism for density map generation is developed. The proposed fully convolutional regressor module adapts UNet with dilated convolution filters to systematically aggregate multi-scale contextual information for density map generation. An attention mechanism represented by prior knowledge of bounding boxes generated by Faster R-CNN is incorporated into the regressor model. This attention mechanism guides the regressor model on where to look for the acne lesions by locating the most salient features related to the understudied acne lesions, therefore improving its robustness to diverse facial acne lesion distributions in sparse and dense regions. Finally, integrating over the generated density maps yields the count of acne lesions within an image, and subsequently, the acne count indicates the level of acne severity. The obtained results demonstrated improved performance compared to the state-of-the-art methods in terms of regression and classification metrics.

The second skin lesion analysis method was considered to study melanoma-related lesions from dermoscopic images using classification-based deep learning methods. Melanoma is the most invasive skin cancer with the highest risk of death. While it is serious skin cancer, it is highly curable if detected early. Melanoma diagnosis is difficult, even for experienced dermatologists, due to the wide range of morphologies in skin lesions. Given the rapid development of deep learning algorithms for

melanoma diagnosis, it is crucial to validate and benchmark these models, which is the main challenge of this work. Thus, the research presented a new benchmarking and selection approach based on the multi-criteria analysis method (MCDM), which integrates entropy and Preference Ranking Organization Method for Enrichment of Evaluations (PROMETHEE) methods. The experimental study was carried out in four phases. Firstly, nineteen convolution neural networks (CNNs) were trained and evaluated on a public dataset of 991 dermoscopic images. Secondly, to obtain the decision matrix, ten criteria, including accuracy, classification error, precision, sensitivity, specificity, F1-score, false-positive rate, false-negative rate, Matthews Correlation Coefficient (MCC), and the number of parameters were established. Third, entropy and PROMETHEE methods were integrated to determine the weights of criteria and rank the models. Fourth, the proposed benchmarking framework was validated using the VIKOR method. The obtained results revealed that the ResNet101 model is selected as the optimal diagnosis model for melanoma in the case study data.

Finally, another approach was presented to study melanoma-related lesions from not only dermoscopic images but also clinical images using the classification-based deep learning method. Reliable skin lesion detection is an important prerequisite for the diagnosis of melanoma and other skin diseases. Existing melanoma assessment models consider either pattern analysis methods or seven-point checklist criteria to investigate skin lesions. However, automatic and accurate detection of the skin lesion and, subsequently, melanoma diagnosis remain an unresolved challenge. Furthermore, there are limitations in both approaches and a trade-off between the two assessment strategies. Thus, the research proposed a pattern analysis method incorporated with a seven-point checklist exploiting a convolutional neural network for melanoma diagnosis where the lesion features are extracted automatically. The benefit of features learned automatically from the dermoscopic images through the stacked layers of convolution filters was designed, realised and evaluated. Both clinical and dermoscopic images were considered as input to the developed multiple-input convolutional neural networks (CNNs), where a separate feature extraction model is implemented for each image type. The features

produced from both models are concatenated for interpretation and final lesion-type prediction. The sum of the weight for predicated lesions, which was calculated according to seven-point checklist criteria and then passed into a threshold model to decide whether the image is normal or abnormal (melanoma or non-melanoma). The performance of the developed algorithm is assessed using a dataset of 2000 dermoscopic and clinical images. The results obtained from the proposed system show a convincing and promising ability for lesion detection and automated melanoma diagnosis from dermoscopy and clinical images.

6.2 Summary of Main Findings

In addition to the details of the research contribution presented in Section 1.5 in Chapter 1, this section summarises the main findings of the research theme presented in this thesis as follows:

- The developed computer-based diagnosis tool would greatly benefit and support automated acne lesion severity grading, significantly reducing the manual assessment and evaluation workload.
- The presented benchmarking framework was proven to be useful in exposing the optimal melanoma diagnosis model targeting to ease the selection process of the proper convolutional neural network architecture.
- The findings of conducted evaluation and investigations on various CNN models for melanoma detection would aid and expedite the deployment of artificial intelligence (AI) assisted CAD systems to clinics and hospitals with regard to ease model selection under different criteria.
- Learning the proposed CNN model using clinical features, i.e. 7-point checklist criteria, and automatic features extracted by convolutional filters in the CNN

on difficult and non-standardised images (clinical images) may aid with leveraging the reliability of melanoma diagnosis systems by exposing the model to the challenging data image examples and complex clinical features.

6.3 Limitations and Future Work

The limitations of the presented research and possible future research directions are described as follows:

- Lack of large labelled dataset: it is suggested to implement and train the developed models for acne severity assessment within weakly-supervised or semi-supervised frameworks, pushing forward to weakly supervised learning fashion due to the unavailability of large amounts of annotated data within the medical domain and the fact that partial annotations are more common.

- Deployment of optimal model: other image modalities, such as non-dermoscopic (clinical) images, can also be used to train and test the network architecture of the pre-trained models developed for melanoma detection. It is also recommended to expand the number of training samples and investigate other untested deep-learning training methodologies.

- Multi-task learning scheme: extend experiments conducted in Chapter 5 by developing a model composed of a common feature-pool providing task-shared features for eight tasks. These task-shared features are integrated with task-specific models designed for each individual task, allowing for learning from both types of features (specific and shared) and providing a suitable balance among those tasks.

- Salient clinical features: detecting the area of interest of lesions by localising the seven lesion attributes would help the CNN extract the salient lesion feature from images. Developing a framework composed of a multi-stage learning scheme

for salient feature localisation and then feature extraction and classification could produce an improved system for melanoma diagnosis.

- Validation: a clinical evaluation would need to be conducted in order to confirm the effectiveness of the deep learning techniques described in this thesis. This would make it possible to assess how well the models might work in a clinical setting. Additionally, varying the quality of the data and the number of patients in various clinical contexts would assess the algorithms' robustness.

Bibliography

[1] M. Vatandoost and S. Litkouhi, "The future of healthcare facilities: How technology and medical advances may shape hospitals of the future," *Hospital Practices and Research*, vol. 4, no. 1, pp. 1–11, 2019.

[2] "Skin lesions: Types and when to see a doctor." `https://www.medicalnewstoday.com/articles/skin-lesions#types-and-causes`. Accessed: 2022-05-15.

[3] G. Argenziano and H. P. Soyer, "Dermoscopy of pigmented skin lesions–a valuable tool for early," *The Lancet Oncology*, vol. 2, no. 7, pp. 443–449, 2001.

[4] M. Vestergaard, P. Macaskill, P. Holt, and S. Menzies, "Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting," *British Journal of Dermatology*, vol. 159, no. 3, pp. 669–676, 2008.

[5] L. E. Barnes, M. M. Levender, A. B. Fleischer, and S. R. Feldman, "Quality of life measures for acne patients," *Dermatologic Clinics*, vol. 30, no. 2, pp. 293–300, 2012.

[6] G. Goodman, "Acne and acne scarring: the case for active and early intervention," *Australian Family Physician*, vol. 35, no. 7, 2006.

[7] T. Agnew, G. Furber, M. Leach, and L. Segal, "A comprehensive critique and review of published measures of acne severity," *The Journal of Clinical and Aesthetic Dermatology*, vol. 9, no. 7, p. 40, 2016.

[8] L. A. Ries, D. Harkins, M. Krapcho, A. Mariotto, B. Miller, E. J. Feuer, L. X. Clegg, M. Eisner, M.-J. Horner, N. Howlader, *et al.*, "Seer cancer statistics review, 1975-2003." https://seer.cancer.gov/archive/csr/1975_2003/, 2006. Accessed: 2019-03-15.

[9] Y. Li and L. Shen, "Skin lesion analysis towards melanoma detection using deep learning network," *Sensors*, vol. 18, no. 2, p. 556, 2018.

[10] G. Argenziano, G. Fabbrocini, P. Carli, V. De Giorgi, E. Sammarco, and M. Delfino, "Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the abcd rule of dermatoscopy and a new 7-point checklist based on pattern analysis," *Archives of Dermatology*, vol. 134, no. 12, pp. 1563–1570, 1998.

[11] "Skin cancers." https://newenglandent.com/skin-cancers/. Accessed: 2022-07-31.

[12] A. Hernández Castillo, "Skin lesions: What are they, types, causes, diagnosis, treatment, and more." https://www.osmosis.org/answers/skin-lesions. Accessed: 2022-07-04.

[13] "Skin lesions." https://www.topdoctors.co.uk/medical-dictionary/skin-lesions. Accessed: 2022-07-04.

[14] H. Kimberly, "What's causing this skin lesion?." https://www.healthline.com/health/skin-lesions. Accessed: 2022-07-04.

[15] "Skin cancer statistics." https://www.nhs.uk/conditions/non-melanoma-skin-cancer/. Accessed: 2019-03-15.

[16] "Non-melanoma skin cancer statistics." https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/non-melanoma-skin-cancer. Accessed: 2019-03-15.

[17] "Melanoma skin cancer statistics." https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/melanoma-skin-cancer. Accessed: 2019-03-15.

[18] M. A. Davies, P. Liu, S. McIntyre, K. B. Kim, N. Papadopoulos, W.-J. Hwu, P. Hwu, and A. Bedikian, "Prognostic factors for survival in melanoma patients with brain metastases," *Cancer*, vol. 117, no. 8, pp. 1687–1696, 2011.

[19] J. Feng, N. Isern, S. Burton, and J. Hu, "Studies of secondary melanoma on c57bl/6j mouse liver using 1h nmr metabolomics," *Metabolites*, vol. 3, no. 4, pp. 1011–1035, 2013.

[20] I. Aslam, A. Fleischer, and S. Feldman, "Emerging drugs for the treatment of acne," *Expert Opinion on Emerging Drugs*, vol. 20, no. 1, pp. 91–101, 2015.

[21] B. Adityan, R. Kumari, D. M. Thappa, *et al.*, "Scoring systems in acne vulgaris," *Indian Journal of Dermatology, Venereology, and Leprology*, vol. 75, no. 3, p. 323, 2009.

[22] G. Stephanie, "Causes of skin lesions." https://www.webmd.com/skin-problems-and-treatments/ss/slideshow-skin-lesion-causes. Accessed: 2022-07-04.

[23] "The truth about acne." https://www.drdivyasharma.com/the-truth-about-acne/. Accessed: 2022-07-31.

[24] N. Aspres, I. B. Egerton, A. C. Lim, and S. P. Shumack, "Imaging the skin," *Australasian Journal of Dermatology*, vol. 44, no. 1, pp. 19–27, 2003.

[25] "Non-invasive imaging techniques for diagnosis." https://skincancer.net/procedures-tests/imaging/non-invasive-techniques. Accessed: 2022-07-31.

[26] "Optical imaging." https://www.nibib.nih.gov/science-education/science-topics/optical-imaging. Accessed: 2022-07-31.

[27] J. Malvehy and G. Pellacani, "Dermoscopy, confocal microscopy and other non-invasive tools for the diagnosis of non-melanoma skin cancers and other skin conditions," *Acta Dermato-Venereologica*, 2017.

[28] C. Rosendahl, P. Tschandl, A. Cameron, and H. Kittler, "Diagnostic accuracy of dermatoscopy for melanocytic and nonmelanocytic pigmented lesions," *Journal of the American Academy of Dermatology*, vol. 64, no. 6, pp. 1068–1073, 2011.

[29] M. Goyal, T. Knackstedt, S. Yan, and S. Hassanpour, "Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities," *Computers in Biology and Medicine*, vol. 127, p. 104065, 2020.

[30] Y. Fujisawa, Y. Otomo, Y. Ogata, Y. Nakamura, R. Fujita, Y. Ishitsuka, R. Watanabe, N. Okiyama, K. Ohara, and M. Fujimoto, "Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis," *British Journal of Dermatology*, vol. 180, no. 2, pp. 373–381, 2019.

[31] D. Shitara, P. Ishioka, Y. Alonso Pinedo, L. Palacios Bejarano, C. Carrera Álvarez, J. Malvehy, and S. Puig i Sardà, "Shiny white streaks: a sign of malignancy at dermoscopy of pigmented skin lesions," *Acta Dermato-Venereologica, 2013, vol. 94, num. 2, p. 132-137*, 2014.

[32] B. Nirmal *et al.*, "Dermatoscopy: Physics and principles," *Indian Journal of Dermatopathology and Diagnostic Dermatology*, vol. 4, no. 2, p. 27, 2017.

[33] Y. Pan, D. S. Gareau, A. Scope, M. Rajadhyaksha, N. A. Mullani, and A. A. Marghoob, "Polarized and nonpolarized dermoscopy: the explanation for the

observed differences," *Archives of Dermatology*, vol. 144, no. 6, pp. 828–829, 2008.

[34] D. Smith and T. Bowden, "Using the abcde approach to assess the deteriorating patient," *Nursing Standard (2014+)*, vol. 32, no. 14, p. 51, 2017.

[35] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2021.," *CA: a Cancer Journal for Clinicians*, vol. 71, no. 1, pp. 7–33, 2021.

[36] R. P. Braun, H. S. Rabinovitz, M. Oliviero, A. W. Kopf, and J.-H. Saurat, "Dermoscopy of pigmented skin lesions," *Journal of the American Academy of Dermatology*, vol. 52, no. 1, pp. 109–121, 2005.

[37] J. A. Witkowski and L. C. Parish, "The assessment of acne: an evaluation of grading and lesion counting in the measurement of acne," *Clinics in Dermatology*, vol. 22, no. 5, pp. 394–397, 2004.

[38] B. M. BURKE and W. Cunliffe, "The assessment of acne vulgaris—the leeds technique," *British Journal of Dermatology*, vol. 111, no. 1, pp. 83–92, 1984.

[39] N. Hayashi, H. Akamatsu, M. Kawashima, and A. S. Group, "Establishment of grading criteria for acne severity," *The Journal of Dermatology*, vol. 35, no. 5, pp. 255–260, 2008.

[40] T. Panch, P. Szolovits, and R. Atun, "Artificial intelligence, machine learning and health systems," *Journal of Global Health*, vol. 8, no. 2, 2018.

[41] "Sophia." `https://www.hansonrobotics.com/sophia/`. Accessed: 2023-01-20.

[42] E. R. Davies, *Computer and machine vision: theory, algorithms, practicalities*. Academic Press, 2012.

[43] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep

learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[44] N. Tyagi, "Six major branches of artificial intelligence (AI)." `https://www.analyticssteps.com/blogs/6-major-branches-artificial-intelligence-ai`, 2021. Accessed: 2022-08-09.

[45] I. Arel, D. C. Rose, and T. P. Karnowski, "Deep machine learning-a new frontier in artificial intelligence research [research frontier]," *IEEE Computational Intelligence Magazine*, vol. 5, no. 4, pp. 13–18, 2010.

[46] F. Hu, G.-S. Xia, Z. Wang, X. Huang, L. Zhang, and H. Sun, "Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 5, 2015.

[47] K. Suzuki, "Overview of deep learning in medical imaging," *Radiological Physics and Technology*, vol. 10, no. 3, pp. 257–273, 2017.

[48] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[49] L. R. Medsker and L. Jain, "Recurrent neural networks," *Design and Applications*, vol. 5, pp. 64–67, 2001.

[50] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[51] G. E. Hinton, "Deep belief networks," *Scholarpedia*, vol. 4, no. 5, p. 5947, 2009.

[52] A. Fischer and C. Igel, "An introduction to restricted boltzmann machines," in *Iberoamerican Congress on Pattern Recognition*, pp. 14–36, Springer, 2012.

[53] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[54] N. Codella, J. Cai, M. Abedini, R. Garnavi, A. Halpern, and J. R. Smith, "Deep learning, sparse coding, and svm for melanoma recognition in dermoscopy images," in *International Workshop on Machine Learning in Medical Imaging*, pp. 118–126, Springer, 2015.

[55] B. Al-Bander, A. Mathew, L. Magerand, E. Trucco, and L. Manfredi, "Real-time lumen detection for autonomous colonoscopy," in *Imaging Systems for GI Endoscopy, and Graphs in Biomedical Image Analysis: First MICCAI Workshop, ISGIE 2022, and Fourth MICCAI Workshop, GRAIL 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings*, pp. 35–44, Springer, 2022.

[56] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," *Journal of Digital Imaging*, vol. 32, pp. 582–596, 2019.

[57] X. Chen, X. Wang, K. Zhang, K.-M. Fung, T. C. Thai, K. Moore, R. S. Mannel, H. Liu, B. Zheng, and Y. Qiu, "Recent advances and clinical applications of deep learning in medical image analysis," *Medical Image Analysis*, p. 102444, 2022.

[58] S. Suganyadevi, V. Seethalakshmi, and K. Balasamy, "A review on deep learning in medical image analysis," *International Journal of Multimedia Information Retrieval*, vol. 11, no. 1, pp. 19–38, 2022.

[59] A. Ciptadi, "what deep learning and how it different machine learning?." `https://idm.net.au/article/`

0012488-what-deep-learning-and-how-it-different-machine-learnin 2019. Accessed: 2022-08-09.

[60] D.-X. Zhou, "Theory of deep convolutional neural networks: Downsampling," *Neural Networks*, vol. 124, pp. 319–327, 2020.

[61] D. Palaz, M. Magimai-Doss, and R. Collobert, "End-to-end acoustic modeling using convolutional neural networks for hmm-based automatic speech recognition," *Speech Communication*, vol. 108, pp. 15–32, 2019.

[62] W. Fang, P. E. Love, H. Luo, and L. Ding, "Computer vision for behaviour-based safety in construction: A review and future directions," *Advanced Engineering Informatics*, vol. 43, p. 100980, 2020.

[63] H.-C. Li, Z.-Y. Deng, and H.-H. Chiang, "Lightweight and resource-constrained learning network for face recognition with performance optimization," *Sensors*, vol. 20, no. 21, p. 6114, 2020.

[64] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of Physiology*, vol. 160, no. 1, p. 106, 1962.

[65] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT Press, 2016.

[66] A. Biswal, "Top ten deep learning algorithms you should know in 2022." https://www.simplilearn.com/tutorials/deep-learning-tutorial/deep-learning-algorithm, 2022. Accessed: 2022-08-09.

[67] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

[68] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[69] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, 2017.

[70] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

[71] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks.," in *CVPR*, vol. 1, p. 3, 2017.

[72] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[73] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[74] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1251–1258, 2017.

[75] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6848–6856, 2018.

[76] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8697–8710, 2018.

[77] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7263–7271, 2017.

[78] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*, pp. 6105–6114, PMLR, 2019.

[79] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[80] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[81] Mathworks, "Pretrained deep neural networks." `https://uk.mathworks.com/help/deeplearning/ug/pretrained-convolutional-neural-networks.html#mw_6ab30f15-4136-46b0-b5db-4f9b3b1cfcd2`, 2022. Accessed: 2022-08-10.

[82] Kaggle, "Platform for predictive modelling and analytics competitions." `https://www.kaggle.com/`. Accessed: 2022-08-10.

[83] N. Elgendy and A. Elragal, "Big data analytics: a literature review paper," in *Industrial Conference on Data Mining*, pp. 214–227, Springer, 2014.

[84] ImageNet, "Image classification on imagenet." `https://paperswithcode.com/sota/image-classification-on-imagenet`, 2022. Accessed: 2022-08-30.

[85] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep

learning: Concepts, cnn architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, no. 1, pp. 1–74, 2021.

[86] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[87] M. Stewart, "Simple introduction to convolutional neural networks." https://towardsdatascience.com/simple-introduction-to-convolutional-neural-networks-cdf8d3077b 2019. Accessed: 2022-08-10.

[88] B. Al-Bander, *Retinal Image Analysis Based on Deep Learning*. PhD thesis, Dept. of Electrical Engineering and Electronics, The University of Liverpool (United Kingdom), 2018.

[89] H. Yingge, I. Ali, and K.-Y. Lee, "Deep neural networks on chip-a survey," in *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 589–592, IEEE, 2020.

[90] Stanford, "Cs231n: Convolutional neural networks for visual recognition." https://cs231n.github.io/convolutional-networks/. Accessed: 2022-08-30.

[91] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 315–323, 2011.

[92] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256, 2010.

[93] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015.

[94] D. E. Rumelhart, G. E. Hinton, R. J. Williams, *et al.*, "Learning representations by back-propagating errors," *Cognitive Modeling*, vol. 5, no. 3, p. 1, 1988.

[95] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[96] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.

[97] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[98] G. Hinton, N. Srivastava, and K. Swersky, "Lecture 6a overview of mini–batch gradient descent," *Coursera Lecture Slides https://class. coursera. org/neuralnets-2012-001/lecture, Online*, 2012.

[99] H. P. Soyer, R. Hofmann-Wellenhof, R. H. Johr, *et al.*, *Color atlas of melanocytic lesions of the skin*. Springer Science & Business Media, 2007.

[100] M. Fiorese, E. Peserico, and A. Silletti, "Virtualshave: automated hair removal from digital dermatoscopic images," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5145–5148, IEEE, 2011.

[101] H. Zhou, M. Chen, R. Gass, J. M. Rehg, L. Ferris, J. Ho, and L. Drogowski, "Feature-preserving artifact removal from dermoscopy images," in *Medical Imaging 2008: Image Processing*, vol. 6914, pp. 439–447, SPIE, 2008.

[102] Q. Abbas, M. E. Celebi, and I. F. García, "Hair removal methods: A comparative study for dermoscopy images," *Biomedical Signal Processing and Control*, vol. 6, no. 4, pp. 395–404, 2011.

[103] M. E. Celebi, H. Iyatomi, and G. Schaefer, "Contrast enhancement in dermoscopy images by maximizing a histogram bimodality measure," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, pp. 2601–2604, IEEE, 2009.

[104] K.-A. Norton, H. Iyatomi, M. E. Celebi, S. Ishizaki, M. Sawada, R. Suzaki, K. Kobayashi, M. Tanaka, and K. Ogawa, "Three-phase general border detection method for dermoscopy images using non-uniform illumination correction," *Skin Research and Technology*, vol. 18, no. 3, pp. 290–300, 2012.

[105] H. Castillejos, V. Ponomaryov, L. Nino-de Rivera, and V. Golikov, "Wavelet transform fuzzy algorithms for dermoscopic image segmentation," *Computational and Mathematical Methods in Medicine*, vol. 2012, 2012.

[106] A. Wong, J. Scharcanski, and P. Fieguth, "Automatic skin lesion segmentation via iterative stochastic region merging," *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 6, pp. 929–936, 2011.

[107] X. Yuan, N. Situ, and G. Zouridakis, "Automatic segmentation of skin lesion images using evolution strategies," *Biomedical Signal Processing and Control*, vol. 3, no. 3, pp. 220–228, 2008.

[108] M. Emre Celebi, Q. Wen, S. Hwang, H. Iyatomi, and G. Schaefer, "Lesion border detection in dermoscopy images using ensembles of thresholding methods," *Skin Research and Technology*, vol. 19, no. 1, pp. e252–e258, 2013.

[109] Z. Ma and J. M. R. Tavares, "A novel approach to segment skin lesions in dermoscopic images based on a deformable model," *IEEE journal of biomedical and health informatics*, vol. 20, no. 2, pp. 615–623, 2016.

[110] F. Nachbar, W. Stolz, T. Merkle, A. B. Cognetta, T. Vogt, M. Landthaler, P. Bilek, O. Braun-Falco, and G. Plewig, "The abcd rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions," *Journal of the American Academy of Dermatology*, vol. 30, no. 4, pp. 551–559, 1994.

[111] F. M. Walter, A. T. Prevost, J. Vasconcelos, P. N. Hall, N. P. Burrows, H. C. Morris, A. L. Kinmonth, and J. D. Emery, "Using the 7-point checklist as a diagnostic aid for pigmented skin lesions in general practice: a diagnostic validation study," *British Journal of General Practice*, vol. 63, no. 610, pp. e345–e353, 2013.

[112] J. S. Henning, S. W. Dusza, S. Q. Wang, A. A. Marghoob, H. S. Rabinovitz, D. Polsky, and A. W. Kopf, "The cash (color, architecture, symmetry, and homogeneity) algorithm for dermoscopy," *Journal of the American Academy of Dermatology*, vol. 56, no. 1, pp. 45–52, 2007.

[113] R. Kasmi and K. Mokrani, "Classification of malignant melanoma and benign skin lesions: implementation of automatic abcd rule," *IET Image Processing*, vol. 10, no. 6, pp. 448–455, 2016.

[114] I. Maglogiannis and K. K. Delibasis, "Enhancing classification accuracy utilizing globules and dots features in digital dermoscopy," *Computer Methods and Programs in Biomedicine*, vol. 118, no. 2, pp. 124–133, 2015.

[115] C. Barata, J. S. Marques, and J. Rozeira, "Evaluation of color based keypoints and features for the classification of melanomas using the bag-of-features model," in *International Symposium on Visual Computing*, pp. 40–49, Springer, 2013.

[116] R. Braun, O. Gaide, M. Oliviero, A. Kopf, L. French, J.-H. Saurat, and H. Rabinovitz, "The significance of multiple blue-grey dots (granularity) for the dermoscopic diagnosis of melanoma," *British Journal of Dermatology*, vol. 157, no. 5, pp. 907–913, 2007.

[117] M. Burroni, R. Corona, G. Dell'Eva, F. Sera, R. Bono, P. Puddu, R. Perotti, F. Nobile, L. Andreassi, and P. Rubegni, "Melanoma computer-aided diagnosis: reliability and feasibility study," *Clinical Cancer Research*, vol. 10, no. 6, pp. 1881–1886, 2004.

[118] M. Ramezani, A. Karimian, and P. Moallem, "Automatic detection of malignant melanoma using macroscopic images," *Journal of Medical Signals and Sensors*, vol. 4, no. 4, p. 281, 2014.

[119] Y. Zhou and Z. Song, "Melanoma diagnosis with multiple decision trees," in *Computer Vision Techniques for the Diagnosis of Skin Cancer*, pp. 267–282, Springer, 2014.

[120] J. F. Alcón, C. Ciuhu, W. Ten Kate, A. Heinrich, N. Uzunbajakava, G. Krekels, D. Siem, and G. De Haan, "Automatic imaging system with decision support for inspection of pigmented skin lesions and melanoma diagnosis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 1, pp. 14–25, 2009.

[121] S. Dreiseitl, L. Ohno-Machado, H. Kittler, S. Vinterbo, H. Billhardt, and M. Binder, "A comparison of machine learning methods for the diagnosis of pigmented skin lesions," *Journal of Biomedical Informatics*, vol. 34, no. 1, pp. 28–36, 2001.

[122] R. Garnavi, M. Aldeen, and J. Bailey, "Computer-aided diagnosis of melanoma using border-and wavelet-based texture analysis," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 6, pp. 1239–1252, 2012.

[123] M. Binder, H. Kittler, A. Seeber, A. Steiner, H. Pehamberger, and K. Wolff, "Epiluminescence microscopy-based classification of pigmented skin lesions using computerized image analysis and an artificial neural network.," *Melanoma Research*, vol. 8, no. 3, pp. 261–266, 1998.

[124] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, 2016.

[125] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, p. 115, 2017.

[126] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 168–172, IEEE, 2018.

[127] H. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. B. H. Hassen, L. Thomas, A. Enk, *et al.*, "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Annals of Oncology*, vol. 29, no. 8, pp. 1836–1842, 2018.

[128] P. Tschandl, N. Codella, B. N. Akay, G. Argenziano, R. P. Braun, H. Cabo, D. Gutman, A. Halpern, B. Helba, R. Hofmann-Wellenhof, *et al.*, "Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study," *The Lancet Oncology*, vol. 20, no. 7, pp. 938–947, 2019.

[129] P. Tschandl, C. Rosendahl, B. N. Akay, G. Argenziano, A. Blum, R. P. Braun, H. Cabo, J.-Y. Gourhant, J. Kreusch, A. Lallas, *et al.*, "Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks," *JAMA Dermatology*, vol. 155, no. 1, pp. 58–65, 2019.

[130] T. J. Brinker, A. Hekler, A. H. Enk, J. Klode, A. Hauschild, C. Berking, B. Schilling, S. Haferkamp, D. Schadendorf, T. Holland-Letz, *et al.*, "Deep

learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task," *European Journal of Cancer*, vol. 113, pp. 47–54, 2019.

[131] R. C. Maron, M. Weichenthal, J. S. Utikal, A. Hekler, C. Berking, A. Hauschild, A. H. Enk, S. Haferkamp, J. Klode, D. Schadendorf, *et al.*, "Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks," *European Journal of Cancer*, vol. 119, pp. 57–65, 2019.

[132] H. A. Haenssle, C. Fink, F. Toberer, J. Winkler, W. Stolz, T. Deinlein, R. Hofmann-Wellenhof, A. Lallas, S. Emmert, T. Buhl, *et al.*, "Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions," *Annals of Oncology*, vol. 31, no. 1, pp. 137–143, 2020.

[133] P. N. Srinivasu, J. G. SivaSai, M. F. Ijaz, A. K. Bhoi, W. Kim, and J. J. Kang, "Classification of skin disease using deep learning neural networks with mobilenet v2 and lstm," *Sensors*, vol. 21, no. 8, p. 2852, 2021.

[134] B. Mazoure, A. Mazoure, J. Bédard, and V. Makarenkov, "Dunescan: a web server for uncertainty estimation in skin cancer detection with deep neural networks," *Scientific Reports*, vol. 12, no. 1, pp. 1–10, 2022.

[135] Y. Wu, B. Chen, A. Zeng, D. Pan, R. Wang, and S. Zhao, "Skin cancer classification with deep learning: A systematic review," *Frontiers in Oncology*, vol. 12, 2022.

[136] S. Qamar, P. Ahmad, and L. Shen, "Dense encoder-decoder–based architecture for skin lesion segmentation," *Cognitive Computation*, vol. 13, no. 2, pp. 583–594, 2021.

[137] Q. Wang, L. Sun, Y. Wang, M. Zhou, M. Hu, J. Chen, Y. Wen, and Q. Li, "Identification of melanoma from hyperspectral pathology image using 3d convolutional networks," *IEEE Transactions on Medical Imaging*, vol. 40, no. 1, pp. 218–227, 2020.

[138] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, IEEE, 2016.

[139] J. Yang, X. Sun, J. Liang, and P. L. Rosin, "Clinical skin lesion diagnosis using representations inspired by dermatologist criteria," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1258–1266, 2018.

[140] S. S. Han, M. S. Kim, W. Lim, G. H. Park, I. Park, and S. E. Chang, "Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm," *Journal of Investigative Dermatology*, vol. 138, no. 7, pp. 1529–1538, 2018.

[141] T. J. Brinker, A. Hekler, A. H. Enk, J. Klode, A. Hauschild, C. Berking, B. Schilling, S. Haferkamp, D. Schadendorf, S. Fröhling, *et al.*, "A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task," *European Journal of Cancer*, vol. 111, pp. 148–154, 2019.

[142] S. Alzahrani, B. Al-Bander, and W. Al-Nuaimy, "Attention mechanism guided deep regression model for acne severity grading," *Computers*, vol. 11, no. 3, p. 31, 2022.

[143] X. Wu, N. Wen, J. Liang, Y.-K. Lai, D. She, M.-M. Cheng, and J. Yang, "Joint acne image grading and counting via label distribution learning," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 10642–10651, 2019.

[144] B. Dreno and F. Poli, "Epidemiology of acne," *Dermatology*, vol. 206, no. 1, p. 7, 2003.

[145] V. Goulden, G. Stables, and W. Cunliffe, "Prevalence of facial acne in adults," *Journal of the American Academy of Dermatology*, vol. 41, no. 4, pp. 577–580, 1999.

[146] H. C. Williams, R. P. Dellavalle, and S. Garner, "Acne vulgaris," *The Lancet*, vol. 379, no. 9813, pp. 361–372, 2012.

[147] S. Alzahrani, B. Al-Bander, and W. Al-Nuaimy, "A comprehensive evaluation and benchmarking of convolutional neural networks for melanoma diagnosis," *Cancers*, vol. 13, no. 17, 2021.

[148] M. Amini, F. Vasefi, M. Valdebran, K. Huang, H. Zhang, W. Kemp, and N. MacKinnon, "Automated facial acne assessment from smartphone images," in *Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues XVI*, vol. 10497, p. 104970N, International Society for Optics and Photonics, 2018.

[149] M. S. Junayed, A. A. Jeny, S. T. Atik, N. Neehal, A. Karim, S. Azam, and B. Shanmugam, "Acnenet-a deep cnn based classification approach for acne classes," in *2019 12th International Conference on Information & Communication Technology and System (ICTS)*, pp. 203–208, IEEE, 2019.

[150] F. S. Abas, B. Kaffenberger, J. Bikowski, and M. N. Gurcan, "Acne image analysis: lesion localization and classification," in *Medical Imaging 2016: Computer-Aided Diagnosis*, vol. 9785, p. 97850B, International Society for Optics and Photonics, 2016.

[151] X. Shen, J. Zhang, C. Yan, and H. Zhou, "An automatic diagnosis method of facial acne vulgaris based on convolutional neural network," *Scientific Reports*, vol. 8, no. 1, pp. 1–10, 2018.

[152] A. Malik, J. Humayun, N. Kamel, and F.-B. Yap, "Novel techniques for enhancement and segmentation of acne vulgaris lesions," *Skin Research and Technology*, vol. 20, no. 3, pp. 322–331, 2014.

[153] T. Chantharaphaichi, B. Uyyanonvara, C. Sinthanayothin, and A. Nishihara, "Automatic acne detection for medical treatment," in *2015 6th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES)*, pp. 1–6, IEEE, 2015.

[154] N. Alamdari, K. Tavakolian, M. Alhashim, and R. Fazel-Rezai, "Detection and classification of acne lesions in acne patients: A mobile application," in *2016 IEEE International Conference on Electro Information Technology (EIT)*, pp. 0739–0743, IEEE, 2016.

[155] Z. Liu and J. Zerubia, "Towards automatic acne detection using a mrf model with chromophore descriptors," in *21st European Signal Processing Conference (EUSIPCO 2013)*, pp. 1–5, IEEE, 2013.

[156] G. Maroni, M. Ermidoro, F. Previdi, and G. Bigini, "Automated detection, extraction and counting of acne lesions for automatic evaluation and tracking of acne severity," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–6, IEEE, 2017.

[157] K. Min, G.-H. Lee, and S.-W. Lee, "Acnet: Mask-aware attention with dynamic context enhancement for robust acne detection," *arXiv preprint arXiv:2105.14891*, 2021.

[158] A. Melina, N. N. Dinh, B. Tafuri, G. Schipani, S. Nisticò, C. Cosentino, F. Amato, D. Thiboutot, and A. Cherubini, "Artificial intelligence for the objective evaluation of acne investigator global assessment.," *Journal of Drugs in Dermatology: JDD*, vol. 17, no. 9, pp. 1006–1009, 2018.

[159] S. Seité, A. Khammari, M. Benzaquen, D. Moyal, and B. Dréno, "Development and accuracy of an artificial intelligence algorithm for acne grading from smartphone photographs," *Experimental Dermatology*, vol. 28, no. 11, pp. 1252–1257, 2019.

[160] T. Zhao, H. Zhang, and J. Spoelstra, "A computer vision application for assessing facial acne severity from selfie images," *arXiv preprint arXiv:1907.07901*, 2019.

[161] Z. V. Lim, F. Akram, C. P. Ngo, A. A. Winarto, W. Q. Lee, K. Liang, H. H. Oon, S. T. G. Thng, and H. K. Lee, "Automated grading of acne vulgaris by deep learning with convolutional neural networks," *Skin Research and Technology*, vol. 26, no. 2, pp. 187–192, 2020.

[162] R. Ramli, A. S. Malik, A. F. M. Hani, and A. Jamil, "Acne analysis, grading and computational assessment methods: an overview," *Skin Research and Technology*, vol. 18, no. 1, pp. 1–14, 2012.

[163] MedicineWise, "Investigator's global assessment (iga) of acne severity." `https://www.nps.org.au/radar/articles/investigators-global-assessment-iga-of-acne-severity-additional` Accessed: 2020-10-01.

[164] B. Dreno, F. Poli, H. Pawin, C. Beylot, M. Faure, M. Chivot, N. Auffret, D. Moyse, F. Ballanger, and J. Revuz, "Development and evaluation of a global acne severity scale (gea scale) suitable for france and europe," *Journal of the European Academy of Dermatology and Venereology*, vol. 25, no. 1, pp. 43–48, 2011.

[165] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 589–597, 2016.

[166] Y. Li, X. Zhang, and D. Chen, "CSRNET: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1091–1100, 2018.

[167] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 234–241, Springer, 2015.

[168] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[169] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, pp. 91–99, 2015.

[170] W. Xie, J. A. Noble, and A. Zisserman, "Microscopy cell counting and detection with fully convolutional regression networks," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 6, no. 3, pp. 283–292, 2018.

[171] L. Yao, T. Liu, J. Qin, N. Lu, and C. Zhou, "Tree counting with high spatial-resolution satellite imagery based on deep neural networks," *Ecological Indicators*, vol. 125, p. 107591, 2021.

[172] M. S. Norouzzadeh, A. Nguyen, M. Kosmala, A. Swanson, M. S. Palmer, C. Packer, and J. Clune, "Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning," *Proceedings of the National Academy of Sciences*, vol. 115, no. 25, pp. E5716–E5725, 2018.

[173] H. Song, H. Liang, H. Li, Z. Dai, and X. Yun, "Vision-based vehicle detection and counting system using deep learning in highway scenes," *European Transport Research Review*, vol. 11, no. 1, pp. 1–16, 2019.

[174] D. Babu Sam, S. Surya, and R. Venkatesh Babu, "Switching convolutional neural network for crowd counting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5744–5752, 2017.

[175] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1, pp. 886–893, IEEE, 2005.

[176] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[177] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision*, pp. 21–37, Springer, 2016.

[178] L. Boominathan, S. S. Kruthiventi, and R. V. Babu, "Crowdnet: A deep convolutional network for dense crowd counting," in *Proceedings of the 24th ACM International Conference on Multimedia*, pp. 640–644, 2016.

[179] X. Wu, Y. Zheng, H. Ye, W. Hu, T. Ma, J. Yang, and L. He, "Counting crowds with varying densities via adaptive scenario discovery framework," *Neurocomputing*, vol. 397, pp. 127–138, 2020.

[180] M. S. Ibrahim, A. Vahdat, M. Ranjbar, and W. G. Macready, "Semi-supervised semantic image segmentation with self-correcting networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12715–12725, 2020.

[181] R. El Jurdi, C. Petitjean, P. Honeine, and F. Abdallah, "Bb-unet: U-net with bounding box prior," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 6, pp. 1189–1198, 2020.

[182] L. Zhang, M. Shi, and Q. Chen, "Crowd counting via scale-adaptive convolutional neural network," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1113–1121, IEEE, 2018.

[183] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[184] J. Akosa, "Predictive accuracy: A misleading performance measure for highly imbalanced data," in *Proceedings of the SAS Global Forum*, vol. 12, 2017.

[185] V. Lempitsky and A. Zisserman, "Learning to count objects in images," *Advances in Neural Information Processing Systems*, vol. 23, pp. 1324–1332, 2010.

[186] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[187] R. Mehrotra, K. R. Namuduri, and N. Ranganathan, "Gabor filter-based edge detection," *Pattern Recognition*, vol. 25, no. 12, pp. 1479–1494, 1992.

[188] S. Alzahrani, B. Al-Bander, and W. Al-Nuaimy, "A comprehensive evaluation and benchmarking of convolutional neural networks for melanoma diagnosis," *Cancers*, vol. 13, no. 17, p. 4494, 2021.

[189] A. M. Glazer and D. S. Rigel, "Analysis of trends in geographic distribution of us dermatology workforce density," *JAMA Dermatology*, vol. 153, no. 5, pp. 472–473, 2017.

[190] "Skin cancer check-ups a long time coming as australia faces huge shortage of dermatologists." `https://www.abc.net.au/news/2021-06-14/gps-to-help-ease-growing-skin-specialist-waiting-times/100211834`. Accessed: 2021-08-27.

[191] D. Eedy, "Dermatology: a specialty in crisis," *Clinical Medicine*, vol. 15, no. 6, p. 509, 2015.

[192] "Royal college of physicians dermatology. london: RCP." `www.rcplondon.ac.uk/sites/default/files/dermatology.pdf`. Accessed: 2021-08-27.

[193] "British association of dermatologists clinical services. london: BAD." `www.bad.org.uk/healthcare-professionals/clinical-services/`. Accessed: 2021-08-27.

[194] S. Alzahrani, W. Al-Nuaimy, and B. Al-Bander, "Seven-point checklist with convolutional neural networks for melanoma diagnosis," in *2019 8th European Workshop on Visual Information Processing (EUVIP)*, pp. 211–216, IEEE, 2019.

[195] N. Nami, E. Giannini, M. Burroni, M. Fimiani, and P. Rubegni, "Teledermatology: state-of-the-art and future perspectives," *Expert Review of Dermatology*, vol. 7, no. 1, pp. 1–3, 2012.

[196] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[197] A. Naeem, M. S. Farooq, A. Khelifi, and A. Abid, "Malignant melanoma classification using deep learning: datasets, performance measurements, challenges and opportunities," *IEEE Access*, vol. 8, pp. 110575–110597, 2020.

[198] E. Pérez, O. Reyes, and S. Ventura, "Convolutional neural networks for the automatic diagnosis of melanoma: An extensive experimental study," *Medical Image Analysis*, vol. 67, p. 101858, 2021.

[199] R. L. Keeney, H. Raiffa, and R. F. Meyer, *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge University Press, 1993.

[200] R. Baltussen and L. Niessen, "Priority setting of health interventions: the need for multi-criteria decision analysis," *Cost effectiveness and resource allocation*, vol. 4, no. 1, pp. 1–9, 2006.

[201] P. Thokala, N. Devlin, K. Marsh, R. Baltussen, M. Boysen, Z. Kalo, T. Longrenn, F. Mussen, S. Peacock, J. Watkins, *et al.*, "Multiple criteria decision analysis for health care decision making—an introduction: report 1 of the ispor mcda emerging good practices task force," *Value in Health*, vol. 19, no. 1, pp. 1–13, 2016.

[202] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

[203] A. Jahan, K. L. Edwards, and M. Bahraminasab, *Multi-criteria decision analysis for supporting the selection of engineering materials in product design*. Butterworth-Heinemann, 2016.

[204] I. Ivlev, J. Vacek, and P. Kneppo, "Multi-criteria decision analysis for supporting the selection of medical devices under uncertainty," *European Journal of Operational Research*, vol. 247, no. 1, pp. 216–228, 2015.

[205] M. Behzadian, R. B. Kazemzadeh, A. Albadvi, and M. Aghdasi, "Promethee: A comprehensive literature review on methodologies and applications," *European Journal of Operational research*, vol. 200, no. 1, pp. 198–215, 2010.

[206] J. J. Liou, C.-Y. Tsai, R.-H. Lin, and G.-H. Tzeng, "A modified vikor multiple-criteria decision method for improving domestic airlines service quality," *Journal of Air Transport Management*, vol. 17, no. 2, pp. 57–61, 2011.

[207] J. Hainmueller, "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies," *Political Analysis*, pp. 25–46, 2012.

[208] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, IEEE, 2009.

[209] S. Alzahrani and W. Al-Nuaimy, "Deep learning approach for skin lesion attributes detection and melanoma diagnosis," in 2^{nd} *International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI)*, pp. 222–223, IFSA, 2020.

[210] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 538–546, 2019.

[211] H. Mirzaalian, T. K. Lee, and G. Hamarneh, "Learning features for streak detection in dermoscopic color images using localized radial flux of principal intensity curvature," in *2012 IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, pp. 97–101, IEEE, 2012.

[212] A. Madooei, M. S. Drew, M. Sadeghi, and M. S. Atkins, "Automatic detection of blue-white veil by discrete colour matching in dermoscopy images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 453–460, Springer, 2013.

[213] G. Fabbrocini, V. De Vita, S. Cacciapuoti, G. Di Leo, C. Liguori, A. Paolillo, A. Pietrosanto, and P. Sommella, "Automatic diagnosis of melanoma based on the 7-point checklist," in *Computer Vision Techniques for the Diagnosis of Skin Cancer*, pp. 71–107, Springer, 2014.

[214] T. Wadhawan, N. Situ, H. Rui, K. Lancaster, X. Yuan, and G. Zouridakis, "Implementation of the 7-point checklist for melanoma detection on smart handheld devices," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3180–3183, IEEE, 2011.

[215] N. C. Codella, Q.-B. Nguyen, S. Pankanti, D. Gutman, B. Helba, A. Halpern, and J. R. Smith, "Deep learning ensembles for melanoma recognition in dermoscopy images," *IBM Journal of Research and Development*, vol. 61, no. 4, pp. 5–1, 2017.

[216] J. Kawahara, A. BenTaieb, and G. Hamarneh, "Deep features to classify skin lesions," in *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, pp. 1397–1400, IEEE, 2016.

[217] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Transactions on Medical Imaging*, vol. 36, no. 4, pp. 994–1004, 2017.

[218] U.-O. Dorj, K.-K. Lee, J.-Y. Choi, and M. Lee, "The skin cancer classification using deep convolutional neural network," *Multimedia Tools and Applications*, pp. 1–16, 2018.

[219] P. Carli, E. Quercioli, S. Sestini, M. Stante, L. Ricci, G. Brunasso, and V. De Giorgi, "Pattern analysis, not simplified algorithms, is the most reliable method for teaching dermoscopy for melanoma diagnosis to residents in dermatology," *British Journal of Dermatology*, vol. 148, no. 5, pp. 981–984, 2003.

[220] G. Argenziano, C. Catricalà, M. Ardigo, P. Buccini, P. De Simone, L. Eibenschutz, A. Ferrari, G. Mariani, V. Silipo, I. Sperduti, *et al.*, "Seven-point checklist of dermoscopy revisited," *British Journal of Dermatology*, vol. 164, no. 4, pp. 785–790, 2011.

[221] G. Argenziano, H. P. Soyer, S. Chimenti, R. Talamini, R. Corona, F. Sera, M. Binder, L. Cerroni, G. De Rosa, G. Ferrara, *et al.*, "Dermoscopy of pigmented skin lesions: results of a consensus meeting via the internet," *Journal of the American Academy of Dermatology*, vol. 48, no. 5, pp. 679–693, 2003.