August 2022

Novel Deep Neural Network for Medical Image Classification

DM Anisuzzaman
University of Wisconsin-Milwaukee

NOVEL DEEP NEURAL NETWORK FOR MEDICAL IMAGE CLASSIFICATION

by

D. M. Anisuzzaman

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Engineering

at

The University of Wisconsin–Milwaukee

August 2022

ABSTRACT

NOVEL DEEP NEURAL NETWORK FOR MEDICAL IMAGE CLASSIFICATION

by

D. M. Anisuzzaman

The University of Wisconsin–Milwaukee, 2022

Under the Supervision of Professor Zeyun Yu

Medical image classification is an essential part of diagnosis, which with automation may benefit both physicians and patients in terms of time and cost. For automation, different Artificial intelligence (AI) methods, including Machine Learning (ML) and Deep Learning (DL), are used widely. Specifically, DL algorithms have become popular in classifying medical images due to their propensity for good performance. This thesis studies medical image classification problems using deep learning models. Four specific medical applications are considered: (1) Osteosarcoma cancer classification in histological images, (2) Burn wound classification, (3) Wound severity classification from clinical images, and (4) Wound type classification using wound images and their corresponding locations. Alongside these classifications, a pre-processing task of automatic wound region of interest (ROI) detection is also performed using a deep neural network. Transfer learning models are used for osteosarcoma classification due to the scarcity of data, and state-of-the-art performance is achieved. In burn wound and wound severity classifications, transfer learning, end-to-end learning, and stacked deep learning models are used. Both classifications

show promising performance. Finally, a novel deep learning multi-modality model is developed to handle image and categorical modalities. This model takes wound images and their corresponding locations as input and predicts the wound types based on the information from both modalities. State-of-the-art performance is achieved with this developed network. Public datasets are used for osteosarcoma and burn wound classifications. Wound datasets are designed for wound localization, wound severity, and wound type classifications with the supervision of wound specialists. A body map is also developed for wound location labeling.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Prof. Zeyun Yu, my mentor and research supervisor, deserves special thanks for his unwavering support, insightful counsel, and patience during my Ph.D. studies. Your advice and support and your knowledge and constructive criticism have always aided me in dealing with difficulties and overcoming barriers. I feel honored to be a part of Prof. Yu's Big Data Analytics and Visualization Lab. Thank you for providing me with the opportunity to work under your supervision.

I would like to offer my heartfelt gratitude to my thesis committee members: Prof. Sandeep Gopalakrishnan, Prof. Rohit Kate, Prof. Jun Zhang, and Prof. Tian Zhao. Your intelligent remarks, unbiased advice, and challenging questions encouraged me to broaden my study to include a variety of viewpoints.

My gratitude also extends to my lab mates for their help and friendliness.

My family members have huge significance in pursuing this degree. I can't thank them enough for their love, crucial counsel, encouragement, direction, and support throughout my life. I would like to thank my father, mother, father-in-law, mother-in-law, brothers, and sister-in-law for all your help, counseling, and encouragement from the bottom of my heart.

My gorgeous and loving wife, Sumaya, has greatly supported me during this challenging path to my Ph.D. I couldn't have done it without you since you made me stronger in dealing with problems and challenges. Thank you once again for everything.

Chapter 1

Introduction

1.1 Artificial Intelligence in Medical Image Analysis

Medical image analysis using artificial intelligence (AI) methods has immense importance as it could save thousands of lives with an adequately designed computer-aided system (CAD). There are mainly three types of AI methods used in medical image analysis: 1. Rule-based methods, 2. Machine learning methods, and 3. Deep learning methods.

Rule-based algorithms require some manually defined rules upon which decisions can be made. For example, association rules are rule-based machine learning methods depending on if-then statements. Zhao et al. [1] presented a review of rule-based methods used to segment blood vessels. The mentioned methods for this segmentation task are Hessian matrix, marching filtering, mathematical morphology, minimal path, active contour, and graph-based methods. Duryea et al. [2] developed a trainable rule-based algorithm for measuring knee joint space from radiographic images. Osl et al. [3] developed a rule-based algorithm named associative voting (AV) to identify biomarker candidates in prostate cancer. Despite numerous rule-based algorithms available for medical (including wound) data processing, machine learning is a new trend that surpasses these methods in many fields.

Machine learning does not require handcrafted rules; instead, an algorithm or function is learned from the given data. A machine learning algorithm can be supervised or unsupervised. In a supervised machine learning algorithm, the input data and the corresponding output (labels) are given; the algorithm learns itself to map the input data to the corresponding outcome. Some popular

supervised algorithms include Bayesian network (BN), Naïve Bayes (NB), logistic regression (LR), decision tree (DT), random forest (RF), support vector machine (SVM), k-nearest neighbor (KNN), neural networks (NN), discriminant analysis (DA), single- and multi-layered perceptrons (MLP), radial basis function networks (RBF), etc. [3], [4]. On the other hand, unsupervised machine learning works with unlabeled data. Given only input data without corresponding outputs, an unsupervised algorithm learns the data pattern and divides them into different clusters. Some unsupervised machine learning algorithms are Markov random field, Bayesian information criterion (BIC), hierarchical clustering (GDLU, AGDL), spectral clustering, k-means, tree matching, independent component analysis (ICA), principal component analysis (PCA), decision trees, etc. [5]. Giger reviewed machine learning in medical imaging and mentioned linear discriminant analysis, SVM, DT, RF, and NN, as some popular machine learning algorithms in this field [6]. A more recent and powerful machine learning method is known as deep learning, as briefly overviewed below.

As a subset of machine learning inspired by the human brain, deep learning does not require any human-designed rules; instead, it demands a large amount of data to map the input to specific labels (supervised learning) or clusters (unsupervised learning). Voulodimos et al. [7] surveyed deep learning algorithms in computer vision, and they discussed some widely used algorithms: Convolutional Neural Networks, Deep Belief Networks, Deep Boltzmann Machines, and Stacked (Denoising) Autoencoders. In addition, LeNet, AlexNet, VGG 19, GoogleNet, ResNet, FCNN, RNNs, Auto-encoders, Stacked Auto-encoders, Restricted Boltzmann Machines, and Deep Belief Networks, Variational Auto-encoders, and Generative Adversarial Networks have been discussed as popular deep learning methods for medical image analysis [8]. Furthermore, Bakator et al. [9] reviewed Convolutional Neural Networks (CNN), Restricted Boltzmann Machine (RBM), Self-

Advised Support Vector Machine (SA-SVM), Convolutional Recurrent Neural Network (CRNN), Deep Belief Network (DBN), Stacked Denoising Autoencoders (SDAE), Undirected Graph Recursive Neural Networks (UGRNN), U-NET, and Class Structure-Based Deep Convolutional Neural Network (CSDCNN) as deep learning methods in the field of medical diagnosis.

In this thesis, two specific medical image classification problems using deep learning models are considered: (1) Osteosarcoma cancer classification in histological images and (2) Wound classification, which includes (i) wound localization, (ii) burn wound classification, (iii) wound severity classification, and (iv) multimodal wound classification with images and wound locations.

1.2 Osteosarcoma Classification

Primary bone tumors account for 5–10% of all new pediatric cancer diagnoses. Osteosarcoma is the most common form of malignant primary bone tumor under the category of bone tumors. Despite the limited number of approximately 1,000 new cases every year in the United States, the prognosis of osteosarcoma remains a challenging issue [10]. There are two age peaks of incidence among patients, with a peak age of children under age ten and adolescents at age 10–20 [11]. Osteosarcoma cancer usually occurs in the metaphysis of long bones on lower limbs, consisting of 40–50% of the total cases [10]. The symptoms of osteosarcoma usually begin with mild localized bone pain, redness, and warmth at the tumor site. Common symptoms include the patient's increasing pain, which often affects patients' movement and joint functions. The early phase of osteosarcoma, if not treated, often results in a wide range of metastasis to other parts of the body, such as at lungs, other bones, and soft tissues [12].

Histological biopsy, X-ray, and magnetic resonance image consist of essential diagnosis of osteosarcoma. Currently, the diagnosis of osteosarcoma includes an initial detailed medical history taking and physical examinations [13]. The symptoms that may direct osteosarcoma include deep-seated, constant, gnawing pain and swelling at the affected site. Pain in multiple areas may portend skeletal metastasis; therefore, they should be investigated appropriately [13]. Beyond the examination, a further evaluation of potential osteosarcoma includes the following procedures: (1) X-ray of the entire affected bone: It is one of the most common ways to diagnose potential tumors. However, the diagnosis of suspicious tumors often requires further confirmation [12]. (2) Magnetic resonance imaging (MRI) scan of the entire affected bone: Doctors use MRI scans frequently for diagnosing joint and bone problems. MRI creates pictures of soft tissue parts of the body that are sometimes hard to see using other imaging tests [10]. (3) Laboratory test is a percutaneous image-guided biopsy [13]. Other tests can suggest that cancer is present, but a biopsy can make a diagnosis. One drawback is that the preparation of histological specimens is time-consuming. For example, accurate detection of osteosarcoma malignancy requires the preparation of at least 50 histology slides to represent a plane of a large three-dimensional tumor [11].

Due to the rise in cancer incidence and patient-specific treatment options, diagnosis and treatment of cancer are becoming more complex [14]. Pathologists must spend an extremely long time examining a large number of slides; therefore, detecting the nuances of histological images can be difficult [15]. The misdiagnosis often occurs due to the extensive work that decreases the accuracy of diagnosis. The osteoblasts' morphology has little difference in differentiated cells, making the image barely distinguishable. Also, the biopsy is a vital and time-consuming step to determine the presence of malignant tissue. The emergence of digital pathology provides new chances of developing new algorithms and software. A histological image can be quantified in

such a system to improve the pathological procedures. The system digitizes glass slides with stained tissue sections at high-resolution images, making computerized image analysis viable [16]. CAD technology integrates powerful algorithms, such as deep learning to accurately recognize tumor malignancy.

This study applied histological medical image analysis based on transfer learning to the pathology archives at Children's Medical Center dataset [17]. Two modified transfer learning approaches, including VGG 19 and Inception V3 models, were applied to the data. The novelty of the study is applying the models to different categories of the dataset and using the whole tile image as input.

1.3 Wound Classification

More than 8 million people are suffering from wounds, and the Medicare cost related to wound treatments ranged from $28.1 billion to $96.8 billion, according to a 2018 retrospective analysis [18]. This immense number can give us an idea of the population related to wound and their care and management. The most common types of wounds/ulcers are diabetic foot ulcer (DFU), venous leg ulcer (VLU), pressure ulcer (PU), and surgical wound (SW). About 34% of people with diabetes have a lifetime risk of developing a DFU, and more than 50% of diabetic foot ulcers become infected [19]. About 0.15% to 0.3% of people suffer from active VLU worldwide [20]. A pressure ulcer is another significant wound, and 2.5 million people are affected each year [21]. Yearly about 4.5% of people have a surgery that leads to a surgical wound [22].

The above statistics show that wounds have caused a substantial financial burden and may even be life-threatening to patients. Due to the shortage of well-trained wound specialists in primary and rural healthcare settings, many wound patients do not have access to specialized

wound care and updated guidelines. Developments of remote telemedicine systems can significantly benefit patients in distant locations, especially in rural areas, with better diagnostic advice, which becomes more relevant in pandemics like COVID-19 [23]. With increasing uses of artificial intelligence (AI) technologies and portable devices such as smartphones, it is now timely to develop remote and intelligent diagnosis and prognosis systems for wound care. An intelligent system can be highly beneficial for wound care in many ways: improved precision, reduced workload and financial burden, standardized diagnosis and management, and higher quality of patient care [24].

An essential part of this wound care system is wound classification, which involves differentiation among different types of wounds (DFU, VLU, PU, SW, etc.) and wound conditions (infection vs. non-infection, ischemia vs. non-ischemic, etc.); wound severity detection, wound tissue classification, burn depth classification, etc. To prepare proper medication and treatment guidelines, physicians must first detect the correct class of the wound. Until artificial intelligence (AI) advancement, wounds were manually classified by wound specialists. AI can save both time and cost and, in some cases, may give better predictions than humans. In recent years, AI algorithms have evolved into so-called data-driven techniques without human or expert intervention, compared to the early generations of rule-based AI, relying mainly on an expert's knowledge [25].

The wound classification study contains one pre-processing step of wound localization and three types of wound classifications: (1) Wound severity classification from wound images, (2) Burn wound classification from wound images, and (3) Wound type classification from wound images and their corresponding locations.

1.4 Contribution and Goal

The contributions and goals of the osteosarcoma classification are:

(1) We demonstrate that the development of deep learning-based tools can detect osteosarcoma malignancy with high accuracy based on a public dataset. The purpose is to successfully distinguish the typical patterns of the non-tumor, necrotic tumor, and viable tumor with relatively low errors.

(2) To explore a suitable deep learning framework for accurate detection and discover possible features that contribute to performance.

The contributions and goals of the wound classification are:

(1) An automated wound localization model is developed using a deep learning method to select the wound ROIs used for wound classifications.

(2) Wound severity is classified into three categories: red, green, and yellow; where red indicates the most severity, yellow indicates medium severity, and green indicates the lowest severity level.

(3) Burn wound images are classified into binary classes (grafts versus non-graft) and multi-classes (deep dermal, full-thickness, superficial dermal) using deep learning models.

(4) Wound types (diabetic, pressure, surgical, and venous) are classified using wound images and their corresponding locations by a novel multi-modal deep learning model.

The rest of this thesis is organized as follows: chapter 2 discusses the deep learning models used to perform experiments in this thesis, and also the performance metrics used to evaluate their performance; chapter 3 briefly presents all the works related to this thesis; chapter 4 discusses the

dataset, dataset processing, model development, experimental setup, and results of osteosarcoma classification from histological images; chapter 5 focuses on the development of an automated wound localization system using deep learning methods, the dataset used in developing this system, experimental setup, and results; chapter 6 explains two wound classification systems: wound severity classification and burn depth classification, the datasets and methods used in these classifications, the experimental setups and results of both experiments; chapter 7 discusses a novel approach of wound type classification by using both wound images and their corresponding locations on the human body, the datasets and methods used for this experiment, and the results of this classification task. Finally, chapter 8 concludes this thesis and remarks on future directions.

Chapter 2

Deep Learning

Deep learning is a subset of machine learning, essentially a three-layer neural network. These neural networks aim to imitate the activity of the human brain by enabling it to learn from enormous quantities of data. Because they can extract features without any human-designed criteria, deep learning-based approaches have recently become increasingly popular due to the rising amount of annotated data. This thesis performs all the localization and classifications using deep learning models. Several deep neural networks and the performance metrics used to evaluate them are discussed in this chapter.

2.1 Deep Neural Networks

All deep neural networks (DNNs) used in this thesis are briefly discussed here. These DNNs are used for osteosarcoma classification, wound localization, wound image classifications, and multimodality wound classifications. Interested readers should examine the sources provided in each network description to learn more about each network described.

2.1.1 MLP

MLP stands for Multilayer Perceptron [26], which is introduced in 1961. There are at least three-node layers: an input layer, a hidden layer, and an output layer. Each node, with the exception of the input nodes, is a neuron with a nonlinear activation function. Backpropagation is a supervised learning technique used by MLPs during training. MLP is distinguished from a linear

perceptron by its multiple layers and non-linear activation. It can tell the difference between data that isn't linearly separable. MLP is used for wound multimodality classification.

2.1.2 LSTM

LSTM [27] stands for Long Short Tern Memory, first introduced in 1997. LSTMs are a particular type of RNN that can learn long-term dependencies. It is their default behavior to remember information for extended periods of time. Three gates regulate the propagation of activations across time in LSTM cells: the forget gate, the input gate, and the output gate. LSTM is used for wound multimodality classification.

2.1.3 AlexNet

AlexNet is a deep CNN architecture proposed in 2012. It outperformed all classic machine learning techniques [28]. This network is made up of 8 layers, comprising three convolutional layers and two fully connected layers, with a total of 60 million parameters. Convolutional layers make up the first five layers of the network, while fully-connected layers make up the last three. A softmax is linked to the final fully connected layer, which yields 1000 probability values for 1000 class labels. Figure 2.1 shows the AlexNet model architecture. We use this network for wound images, burn wound images, and wound multimodality classifications.

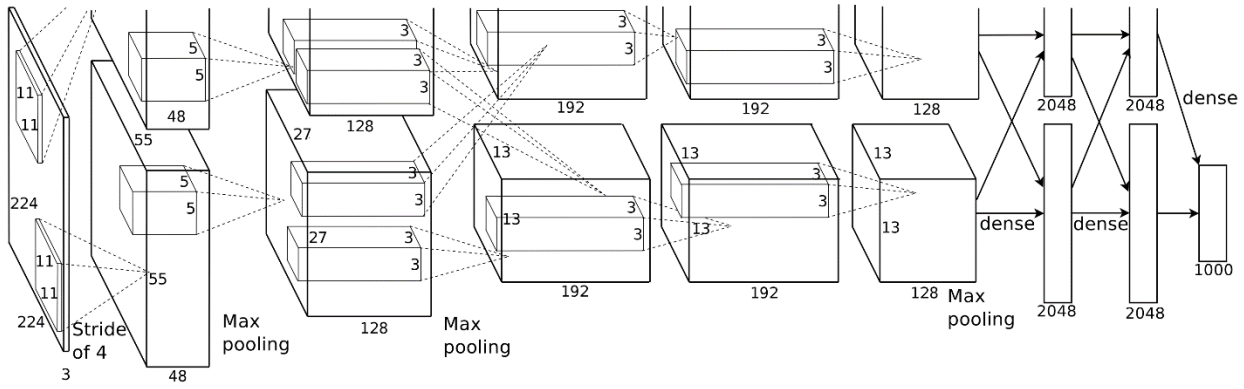


Figure 2.1: AlexNet architecture [28].

2.1.4 VGG16

VGG16 [29] is a deep CNN model proposed in 2014. This network contains five blocks of convolutional layers, each followed by a max-pooling layer. All the convolutional layers use 3 × 3 kernel sizes with increasing filters (depths). The 2 × 2 max-pooling layers are used in all places. The convolutional and max-pooling layers are followed by three fully connected layers. Finally, a "softmax" layer is added for classification. This network is used for osteosarcoma classification, burn wound image, chronic wound image, and multimodality classifications.

2.1.5 VGG19

VGG19 [29] is an extension of VGG16 architecture which is also proposed in 2014. Here, convolutional layers are used for feature extraction, whereas fully connected layers are used for classification. These layers learn a non-linear function between the high-level features given as an output from the previous (convolutional) layers. Figure 2.2 shows the VGG19 model architecture. The figure shows that all the convolution layers use 3 × 3 filters, and all the max-pooling layers use 2 × 2 filters. The FC1 and FC2 layers contain 512 and 1024 neurons, respectively. The softmax

layer's neurons vary depending on the classification task. We use this network for osteosarcoma classification, burn wound image, chronic wound image, and multimodality classifications.

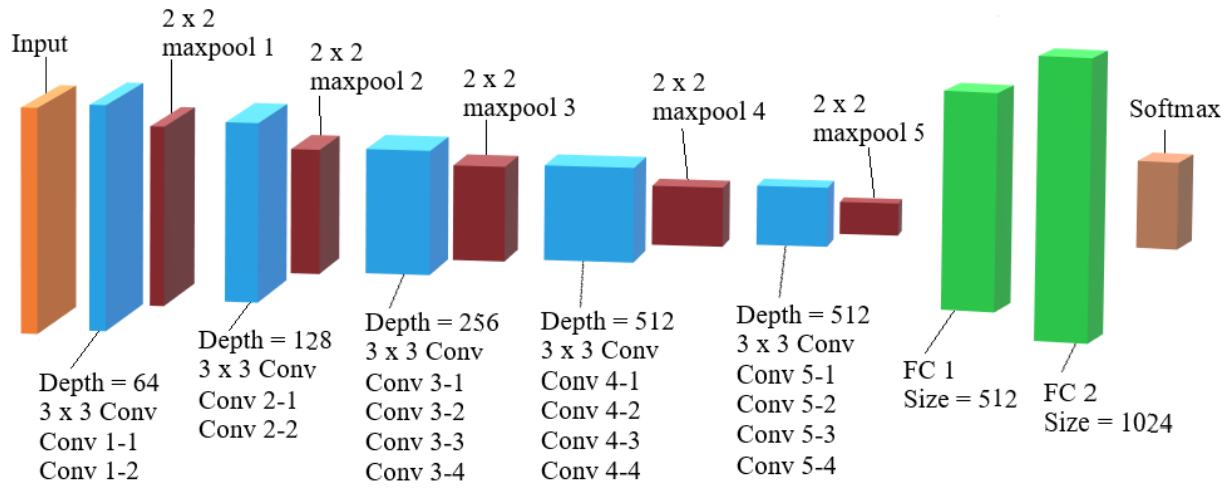


Figure 2.2: VGG19 network architecture.

2.1.6 InceptionV3

Inception V3 [30] network was proposed in 2015. Inception Networks (GoogLeNet) have shown to be more computationally efficient than VGGNet, both in terms of the number of parameters generated and the cost incurred. This network is built upon the following concepts: factorized convolutions, smaller convolutions, asymmetric convolutions, auxiliary classifier, and grid size reduction. Inception V3 network is used for osteosarcoma classification, burn wound image, chronic wound image, and multimodality classifications.

2.1.7 ResNet50

ResNet50 [31], introduced in 2015, is a ResNet variation containing 48 Convolution layers, 1 MaxPool layer, and 1 Average Pool layer. The first convolution layer has a kernel size of 7 × 7 with a depth of 64, followed by a max-pooling layer. This is followed by a group of convolutional blocks (1 × 1, 3 × 3, 1 × 1) with different depths for 48 layers. In each convolutional block, the depth increases from 64, 64, 256 to 512, 512, 2048. Deeper models, in general, yield a higher error, which should not be the case. The authors developed a deep residual learning architecture to address this issue, including shortcut connections that merely conduct identity mappings. The advantage of these shortcut identity mappings is that no more parameters are introduced to the model, and the computational time does not increase. We use ResNet50 for osteosarcoma classification, chronic wound image, and multimodality classifications.

2.1.8 DenseNet201

The DenseNet201 [32] is introduced in 2016, which is a 201-layer deep convolutional neural network. Each layer in DenseNet receives extra input from all preceding layers and sends its own feature maps to all future layers, which are concatenated. So, each layer receives collective knowledge from the layers in front of it. This network has multiple benefits: it solves the vanishing-gradient problem, improves feature propagation, encourages feature reuse, and reduces the number of parameters significantly. DenseNet 201 is used for osteosarcoma classification and chronic wound image classification.

2.1.9 InceptionResNetV2

InceptionResNetV2 [33], introduced in 2016, is a convolutional neural network based on the Inception family but includes residual connections. This 164-layer network incorporates the Inception-ResNet block, which combines several sized convolutional filters with residual connections. The introduction of residual connections solves the deterioration issue caused by deep structures and decreases the training time. In addition, the filter concatenation stage of the Inception architecture is replaced in this network. We use InceptionResNetV2 network for chronic wound image classification.

2.1.10 Xception

Francois Chollet proposed Xception [34] in 2017. It stands for "extreme inception," which pushes Inception's concepts to their logical conclusion. In Inception, 1 × 1 convolutions were used to compress the original input, and different types of filters were applied to each of the depth spaces based on the input spaces. Xception just reverses this process. The feature extraction basis of the Xception architecture is made up of 36 convolutional layers. In summary, the Xception design is a depthwise separable convolution layer stack with residual connections. This network is used for chronic wound image classification.

2.1.11 NasNetLarge

The NASetLarge [35] is designed using Neural Architecture Search (NAS), and inception cells are utilized to build the model's layer. This network is proposed in 2018. Normal cells and reduction cells are the two types of cells utilized to create NASNet. Convolutional cells that return

a feature map of the same dimension are normal cells. Convolutional cells that return a feature map with a two-fold reduction in height and width are called reduction cells. NasNet large is used for osteosarcoma, burn wound, and chronic wound image classifications.

2.1.12 MobileNetV2

MobileNetV2 [36], proposed in 2018, is a convolutional neural network design that aims to be mobile-friendly. It is built on an inverted residual structure, with residual connections between bottleneck layers. As a source of non-linearity, the intermediate expansion layer filters features with lightweight depth-wise convolutions. Overall, the MobileNetV2 architecture includes a fully convolutional layer with 32 filters, followed by 19 residual bottleneck layers. We use this network for chronic wound image classification.

2.1.13 Single Shot MultiBox Detector (SSD)

SSD [37] stands for Single Shot MultiBox Detector, introduced in 2016, composed of mainly two object detection parts: feature maps extraction and object detection by applying convolution filters. VGG16 and Conv4_3 layer is used for feature map extraction and object detection. Each location makes four object predictions, where each prediction consists of a boundary box and scores for each class (including the class for no object), and the highest score is selected as the class for the bounded object. After extracting the feature maps, 3 × 3 convolution filters are applied for each cell to make predictions. Six more auxiliary convolution layers are added after the VGG16; five out of six are used for object detection. SSD makes 8,732 predictions per class from six layers, followed by a non-maximum suppression step to produce the final

detections. The SSD model architecture is shown in Figure 2.3. This network is used for wound localization.

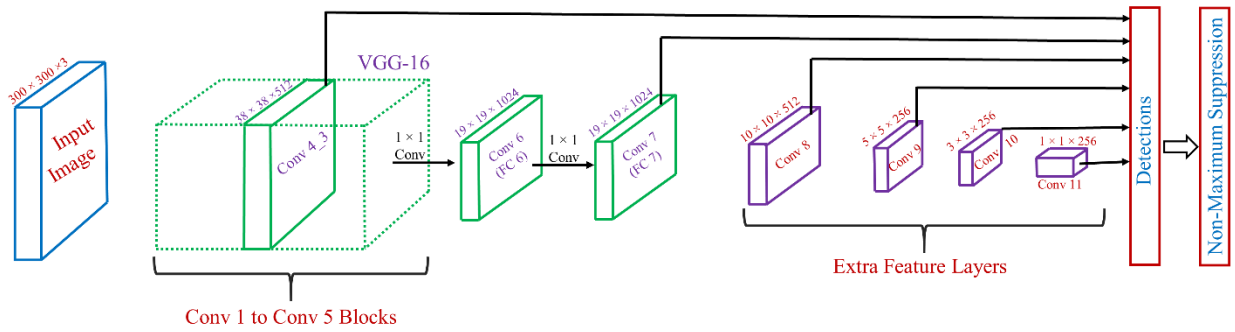


Figure 2.3: SSD network architecture.

2.1.14 You Only Look Once (YOLOv3)

YOLOv3 [38] is the third generation of the YOLO family, introduced in 2018, predicting bounding boxes and classifying the object within the bounding box in one pass. YOLOv3 predicts on a per-frame basis, and no temporal information is employed. This architecture consists of three types of networks: Darknet-53, upsampling, and YOLO layers or detection layers. The darknet-53 network extracts features from the input image, composed of residual blocks as the basic component. Each residual block consists of a pair of 3 × 3 and 1 × 1 convolutional layers together with shortcut connections. As the name suggests, there are 53 convolutional layers in Darknet-53. In the upsampling layers, YOLOv3 has a total of 106 fully convolutional layers. The YOLO layers are responsible for detecting objects at different scales using features extracted by Darknet-53 layers. At the initial YOLO layer, the grid size is 1/32 of the input image size, and at the final YOLO layer, the grid size is 1/8 of the input image size. With three YOLO layers, smaller objects

can also be detected. Each YOLO layer consists of a few convolution layers with batch normalization and leaky ReLU activation. Shortcut connections connect darknet-53 intermediate layers to the layer after upsampling layers. The model architecture of YOLOv3 is shown in Figure 2.4. We used this network for wound localization.

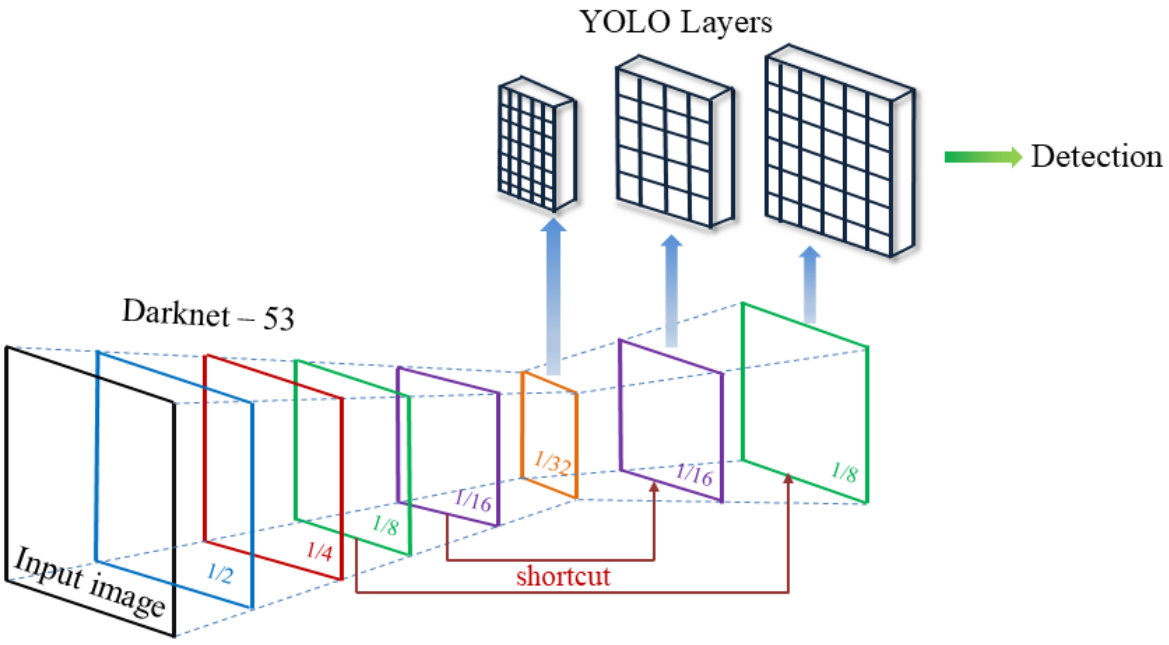


Figure 2.4: YOLOv3 network architecture.

2.2 Performance Metrics

We provide a brief overview and the equations we used to investigate the performance of the localizer and classifiers in this thesis. More details about these performance metrics can be found in [39], [40], and [41].

2.2.1 Intersection over Union (IoU)

Intersection over union measures the overlap between the ground truth box (manual localization) and the predicted box (model result) over their union. The IoU is calculated with Equation (2.1).

$$IoU = \frac{Ground\ Truth\ Box\ \cap\ Predicted\ Box}{Ground\ Truth\ Box\ \cup\ Predicted\ Box} \tag{2.1}$$

2.2.2 Precision, Recall, and F1-score

Precision measures the percentage of correctly classified or localized images in the classification or localization, whereas Recall measures the percentage of correctly classified or localized images in the ground truth. The F1 score is the weighted average of Precision and Recall values. A higher F1 score indicates better performance of the model. Equations (2.2), (2.3), and (2.4) show the definitions of Precision, Recall, and F1-score metrics, respectively.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{2.2}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{2.3}$$

$$F1 - score = 2\ \times \frac{precision\ \times\ recall}{precision + recall} \tag{2.4}$$

2.2.3 Mean Average Precision (mAP)

The mean average Precision compares different object detectors over multiple datasets. mAP value calculation requires interpolated Precision which is simply the highest precision value for a specific recall value. Interpolated Precision is calculated using Equation (2.5). The mAP value is calculated from the summation of interpolated precision values shown in Equation (2.6).

$$P_{interp}(r) = \max_{r' : r' \geq r} p(r') \tag{2.5}$$

$$mAP = \sum_{r=0}^{1} (r_n - r_{n-1}) P_{internp}(r_n) \tag{2.6}$$

2.2.4 Accuracy

Accuracy is the ratio of correctly predicted data to the total amount of data. Equations 2.7 shows the formula for accuracy. In this equation, TP, TN, FP, and FN, represent True Positive, True Negative, False Positive, and False Negative measures.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2.7}$$

2.2.5 ROC and AUC

The receiver operating characteristic (ROC) curve shows the diagnostic ability of a binary classifier system for different thresholds. This curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity). The area under the curve (AUC) indicates that the classifier

gives a randomly chosen positive instance a higher probability than a randomly chosen negative instance.

Chapter 3

Osteosarcoma Classification

3.1 Problem Statement

In the U.S., 5-10% of new pediatric cancer cases are primary bone tumors. The most common type of primary malignant bone tumor is osteosarcoma. In this chapter, we intend to improve the detection and diagnosis of osteosarcoma using computer-aided detection (CAD) and diagnosis (CADx). Such tools as convolutional neural networks (CNNs) can significantly decrease the surgeon's workload and make a better prognosis of patient conditions. However, CNNs need to be trained on a large amount of data to achieve more reliable performance. In this study, transfer learning techniques, pre-trained CNNs, are adapted to a public dataset on osteosarcoma histological images to detect necrotic images from non-necrotic and healthy tissues. First, the dataset was preprocessed, and different classifications were applied. Then, Transfer learning models, including VGG19, and Inception V3 are used and trained on Whole Slide Images (WSI) with no patches to improve the accuracy of the outputs. Finally, the models are applied to different classification problems, including binary and multi-class classifiers. Experimental results show that the accuracy of the VGG19 has the highest, 96%, performance amongst all binary classes and multiclass classification. Our fine-tuned model demonstrates state-of-the-art performance in detecting malignancy of Osteosarcoma based on histologic images.

3.2 Related Works

Deep learning is complex, requiring a more significant number of datasets than traditional machine learning algorithms. However, the lack of histological and radiological images often restricts its development. A typical deep learning model that addresses image classification requires three steps: model building, training on a dataset, and performance evaluation on specific tasks. Knowledge-transferring significantly improves learning outputs if done efficiently while minimizing expensive data labeling efforts. A few studies have already focused on this area in medical image classifications: De Matos et al. [42] used double transfer learning to classify histopathologic images. Noorul Wahab et al. [43] aimed at a more challenging segmentation task and detecting mitotic nuclei. They used a similar hybrid CNN model and achieved a fair AUC value. Other studies include the prediction of pathological invasiveness in lung adenocarcinoma [44], classification of liver cancer histopathology images [45], automated invasive ductal carcinoma detection [46], and skin cancers [47]. Researchers employed deep learning techniques [48]–[50], focusing on the segmentation and classification of histology tissue in tumor image datasets.

The authors [51] reported the first fully automated tool to assess viable and necrotic tumors in osteosarcoma using histological images and deep learning models. The goal is to label the diverse tissue regions as viable tumors, necrotic tumors, and non-tumor. They employed both machine learning and deep learning models. The ensemble learning model achieved an overall accuracy of 93.3%, with class-specific accuracies of 91.9% for non-tumor, 95.3% for viable tumors, and 92.7% for necrotic tumors. A multiple-layered neural network is proficient in image segmentation, as they achieved significantly better performances than machine learning algorithms. In clinical practices, the primary goal is to decrease the mortality of osteosarcoma

diagnosis. In addition, it is imperative to prevent the early-stage tumor from metastasis. Early automatic detection can reduce the chance of misdiagnosis and serve as an assistant tool for the surgeon's preference to determine if metastasis has occurred. Using CNN, adopting computer-aided technology can significantly reduce the surgeon's workload and achieve a better prognosis.

3.3 Methodology

3.3.1 Dataset

The dataset used in the study was obtained from Arunachalam et al. [51]. They provided a data set of osteosarcomas and conducted a variety of machine learning and deep learning techniques. Tumor samples from the Children's Medical Center, Dallas, were collected from the pathology reports of the osteosarcoma resection for 50 patients treated between 1995 and 2015. They selected 40 WSIs of the digitized images representing tumor heterogeneity and response properties in the study. In each WSI, 30 1024 × 1024 pixel image tiles were randomly selected at the 10X magnification factor. In addition, 1,144 of the resulting 1,200 image tiles, such as those that fall into non-fabric, ink marks regions, and blurry images, were chosen after removing irrelevant tiles. Moreover, they generated 56,929 patches of 128 × 128 pixels. Some sample dataset images are shown in Figure 3.1.

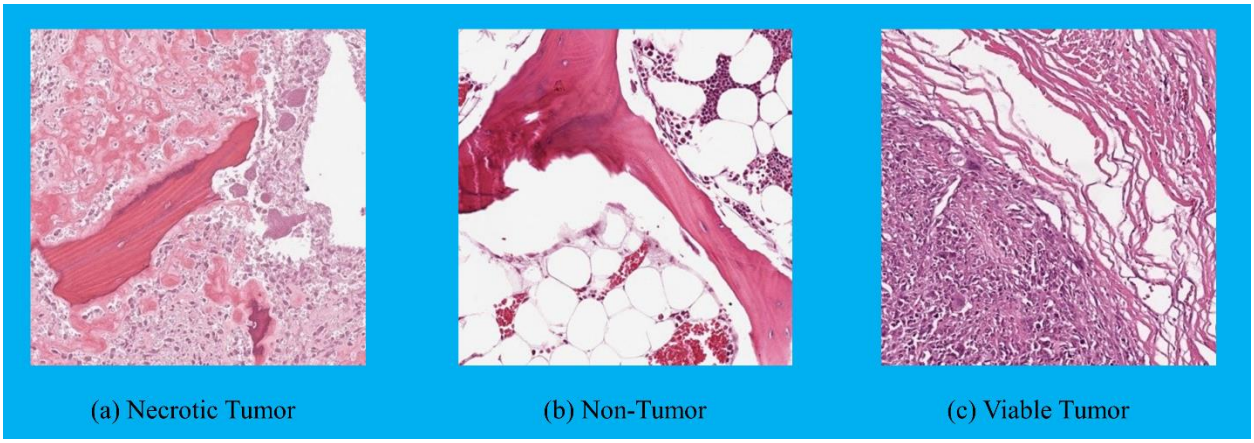


Figure 3.1: Sample images from the osteosarcoma dataset.

3.3.2 Data Preprocessing

Original images of 1024 × 1024 pixels were used for model training, validation, and evaluation. First, we split the datasets into training, validation, and testing images at a ratio of 70%, 10%, and 20%, respectively. The data are then augmented using an image data generator module of "Keras" [52]. In this step, all image intensities are first rescaled to 0 to 1. Next, the following augmentations have been performed: rotation, width shift, height shift, vertical flip, and horizontal flip. Finally, due to memory limitations, we down-sampled the original images by passing the input shape of 375 × 375 rather than 1024 × 1024.

3.3.3 Model Selection

There are 26 deep learning models in Keras Applications that can be used for prediction, feature extraction, and fine-tuning [53]. Six of these models are applied for multi-class classification, and among them, we have chosen the best model for our experiment depending on

the test accuracy. Table 3.1 shows the test results of these models. VGG19 gives the best result among these models, and we choose this model for future experiments.

From Table 3.1, we can see that ResNet50 gives the inferior result among all the models. DenseNet201 gives the second most inferior result. Both networks have very deep layers and complex architecture. With the small number of images in our selected dataset, these networks are underfitting during the training process, which is reflected in the testing results. InceptionV3 and NASNetLarge give almost the same results for all performance metrics (precision, recall, f1-score, and accuracy). Although they show nearly 80% accuracy, they are still far away from the accuracy given by VGG16 and VGG19 networks. VGG16 and VGG19 are the most simple networks among the selected networks, and they both perform pretty well with our selected dataset. As VGG19 is an extension of the VGG16 network, we later choose the InceptionV3 network for comparison with the best model (VGG19) due to their divergent network architecture.

Table 3.1: Multiclass classification results of various models on osteosarcoma dataset.

Model	**Weighted Average Precision**	**Weighted Average Recall**	**Weighted Average F1-Score**	**Accuracy**
VGG16	0.89	0.88	0.88	0.883
VGG19	0.94	0.94	0.94	0.939
ResNet50	0.22	0.47	0.30	0.470
InceptionV3	0.81	0.78	0.79	0.783
DenseNet201	0.61	0.58	0.56	0.583
NASNetLarge	0.80	0.79	0.79	0.791

3.3.4 VGG19 Network Modification

Keras applications are used for importing the VGG19 model. Pre-trained weights have been used for model training. The fully connected layer along with the output layer of the VGG19 model is discarded. Two fully connected layers have been added after the last “maxpool” layer. As the dataset is not very big and a maximum of 3-class classification is performed, adding more fully connected layer(s) does not improve the model performance but increases the training time. Dropout layers are used to avoid over-fitting the training data. We have used “ReLU” (Rectified Linear Unit) activation in the dense layers and the “softmax” activation function in the output layer. ReLU is computationally efficient than the “sigmoid” and “tanh” functions, as it does not need to perform expensive exponential operations. Also, ReLU solves the vanishing gradient problem, as the gradient is either 0 or 1 for this function, and it never saturates, which means the gradients cannot vanish and be transferred perfectly across the network. The “Softmax” activation function is generally used for multiclassification, and its output is a probability distribution, which means the output is mapped to the range of [0,1], and the sum of the total output is 1. The softmax layer’s neurons vary depending on our classification task. For binary and multi-class classification, it includes two and three neurons, respectively.

3.4 Experiments and Results

3.4.1 Setup

We performed four binary classifications and a multiclass (three classes) classification with our dataset containing three classes. In each classification, we applied two models: VGG19 and Inception V3. Inception V3 has been used for model comparison. The models are written in the Python programming language in the Keras deep learning framework. The models are trained and tested on an Nvidia GeForce RTX 2080Ti GPU platform.

The loss functions used for binary and multiclass classifications are binary cross-entropy and categorical cross-entropy, respectively. In both types of classification, the Adam optimizer is applied for minimizing the loss function by updating the weight parameters. The learning rate is set to Keras's default 0.01. Batch size is set to 80, 28, and 16 for training, validation, and testing, respectively. All models are trained for 1500 epochs, with a callback stopping training when validation accuracy exceeds 0.98.

Two-class classifications are evaluated on the following datasets: 1.) Non-Tumor (NT) versus Necrotic Tumor (NCT) and Viable Tumor (VT), 2.) Necrotic Tumor versus Non-Tumor, 3.) Viable Tumor versus Non-Tumor, and 4.) Necrotic Tumor versus Viable Tumor. We also performed the multiclass classification among the NT, NCT, and VT classes. We presented a confusion matrix, precision (equation 2.2), recall (equation 2.3), f1 score (equation 2.4), and accuracy (equation 2.7) for all classifications to evaluate our model performance. We also reported the receiver operating characteristic (ROC) curve and area under the curve (AUC) for all the two-class classifications.

3.4.2 Results

The following sections briefly present the evaluation metrics for all the classifications with two models. Figure 3.2 shows the confusion matrix for all classifications with both networks.

Actual Class	Predicted Class: NT	Predicted Class: NCT+VT
NT	**101**	7
NCT+VT	4	**118**

(a) NT vs. NCT+VT with VGG 19

Actual Class	Predicted Class: NT	Predicted Class: NCT+VT
NT	95	13
NCT+VT	14	108

(b) NT vs. NCT+VT with Inception V3

Actual Class	Predicted Class: NCT	Predicted Class: NT
NCT	**51**	2
NT	5	**103**

(c) NCT vs. NT with VGG 19

Actual Class	Predicted Class: NCT	Predicted Class: NT
NCT	48	5
NT	12	96

(d) NCT vs. NT with Inception V3

Actual Class	Predicted Class: NT	Predicted Class: VT
NT	103	5
VT	3	**66**

(e) NT vs. VT with VGG 19

Actual Class	Predicted Class: NT	Predicted Class: VT
NT	**107**	1
VT	32	37

(f) NT vs. VT with Inception V3

Actual Class	Predicted Class: NCT	Predicted Class: VT
NCT	48	5
VT	4	**65**

(g) NCT vs. VT with VGG 19

Actual Class	Predicted Class: NCT	Predicted Class: VT
NCT	**53**	0
VT	21	48

(h) NCT vs. VT with Inception V3

Actual Class	Predicted Class: NCT	Predicted Class: NT	Predicted Class: VT
NCT	**48**	3	2
NT	2	**103**	3
VT	2	2	**65**

(i) multiclass with VGG 19

Actual Class	Predicted Class: NCT	Predicted Class: NT	Predicted Class: VT
NCT	44	6	3
NT	12	93	3
VT	19	7	43

(j) multiclass with Inception V3

Figure 3.2: Confusion matrixes of all osteosarcoma classifications. Here, NT = Non-Tumor, NCT = Necrotic Tumor, and VT = Viable Tumor.

Tables 3.2 and 3.3 show the precision, recall, and f1 score for all the binary and multiclass classifications with each of the present networks. Figure 3.3 shows the accuracy of the classifiers for all the classifications.

Table 3.2: Precision, Recall, and F1-Score for binary classification on osteosarcoma dataset.

Non-Tumor versus Necrotic Tumor and Viable Tumor						
	Non-Tumor			Necrotic and Viable Tumor		
Networks	Precision	Recall	F1	Precision	Recall	F1
VGG 19	**0.96**	**0.94**	**0.95**	**0.94**	**0.97**	**0.96**
Inception V3	0.87	0.88	0.88	0.89	0.89	0.89
Necrotic Tumor versus Non-Tumor						
	Necrotic Tumor			Non-Tumor		
Networks	Precision	Recall	F1	Precision	Recall	F1
VGG 19	**0.91**	**0.96**	**0.94**	**0.98**	**0.95**	**0.97**
Inception V3	0.80	0.91	0.85	0.95	0.89	0.92
Viable Tumor versus Non-Tumor						
	Non-Tumor			Viable Tumor		
Networks	Precision	Recall	F1	Precision	Recall	F1
VGG 19	**0.97**	0.95	**0.96**	0.93	**0.96**	**0.94**
Inception V3	0.77	**0.99**	0.87	**0.97**	0.54	0.69
Necrotic Tumor versus Viable Tumor						
	Necrotic Tumor			Viable Tumor		
Networks	Precision	Recall	F1	Precision	Recall	F1
VGG 19	**0.92**	0.91	**0.91**	0.93	**0.94**	**0.94**
Inception V3	0.72	**1.00**	0.83	**1.00**	0.70	0.82

Table 3.3: Precision, Recall, and F1-Score for multiclass classification on osteosarcoma dataset.

Multiclass									
	Necrotic Tumor			Non-Tumor			Viable Tumor		
Networks	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
VGG 19	**0.92**	**0.91**	**0.91**	**0.95**	**0.95**	**0.95**	**0.93**	**0.94**	**0.94**
Inception V3	0.59	0.83	0.69	0.88	0.86	0.87	0.88	0.62	0.73

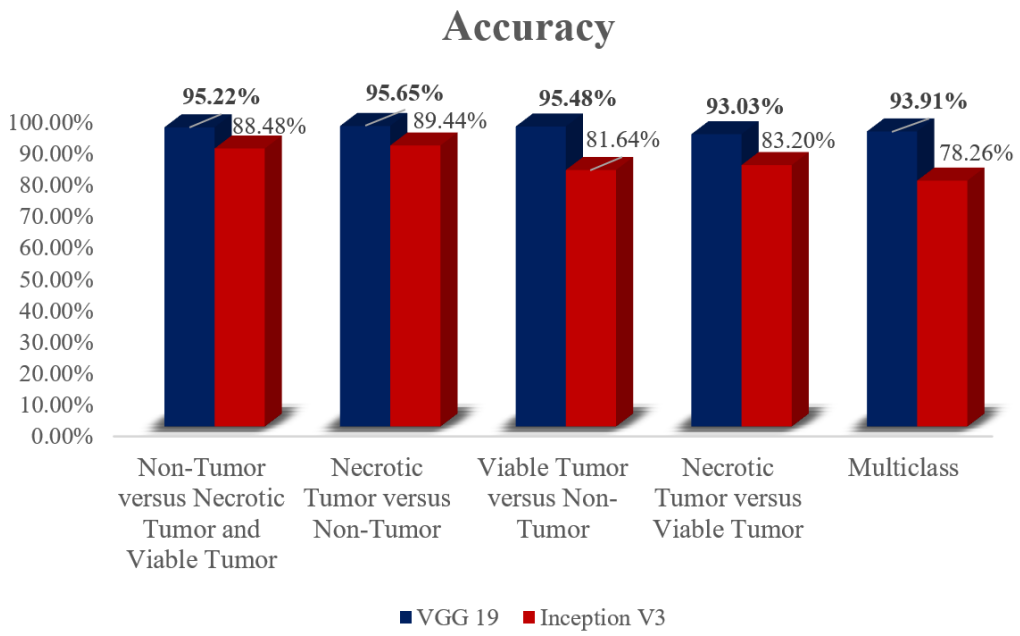


Figure 3.3: Accuracy scores of osteosarcoma classification.

3.5 Discussion

Osteosarcoma is a common tumor in pediatric cases of cancer that requires extensive work of pathologists to confirm the case. While other medical images have already computerized analysis, osteosarcoma histological images are rarely classified using deep learning models. We believe it is possible to use computer-aided technology to help classify and recognize the image of a malignant tumor. This study uses a deep learning-based technique for image classification to detect histologic images and identify osteosarcoma's malignancy. Our study provides the option of using a computer to accelerate diagnosing and detecting osteosarcoma malignancy. Furthermore, we apply and compare two popular network architectures, VGG19 and Inception V3. Thus, we obtain higher performance than prior studies with the same dataset. We have configured and tested models with custom layers to achieve the best performance.

From Figure 3.2, we can see that for NT vs. VT and NCT vs. VT, respectively, the prediction of non-tumor and the necrotic tumor is performed well by Inception V3. However, VGG19 works very robustly in all other cases compared to Inception V3. So, in overall balance, VGG19 beats Inception V3.

From Tables 3.2 and 3.3, we can see that for VT vs. NT and NCT vs. VT cases, the precision of viable tumor and recall of necrotic tumor and nontumor are high for Inception V3. But the interesting fact is that all the f1 scores are higher for the VGG19 model. Since the f1 score indicates the weighted average of precision and recall, a higher f1 score means precision and recall are close to each other for VGG19. For inception V3, only a single metric is higher (either precision or recall), indicating a lower score than the other one. Hence, VGG19 beats inception V3 by a considerable margin in balance in overall performance. From Figure 3.3, for all classifications, VGG19 achieves the highest accuracy.

From Figure 3.4, we can see that VGG19 has the highest AUC value for all binary (two-class) classifications. The AUC values are impressive (0.95, 0.96, 0.96, and 0.92 for non-tumor versus necrotic tumor and viable tumor, necrotic tumor versus non-tumor, viable tumor versus nontumor, and necrotic tumor versus viable tumor classifications, respectively), which assures us with excellent reliability. So, from all the above analytical discussions, it is safe to say that VGG19 works well for all classifications. While Inception V3 has three types of convolutions (1×1, 3×3, 5×5), VGG19 has only one type of convolution (3×3). Instead of going deeper, Inception V3 goes wider on image feature searching. As our dataset contains biopsy images in which some parts may only include some specific features of a particular class (necrotic or viable), some of the inception kernels may not provide good qualities, and in the concatenation layer, the performance may decrease. In VGG19, the kernel size is always the same (3×3), which may lead to better

classification accuracy specifically for our dataset. However, this dataset has a small number of images (1144), which is not suitable for deep learning models. Deep learning demands lots of data to learn the connection between the input and the corresponding output. To overcome the data limitation problem, we applied the transfer learning approach. Both VGG19 and inception V3 are pre-trained with the Imagenet dataset, where all the low-level features (edge, curve, etc.) are trained with the Imagenet dataset, and we transfer that learned weights to our dataset. The fully connected and output layers are replaced in both models and trained with our dataset.

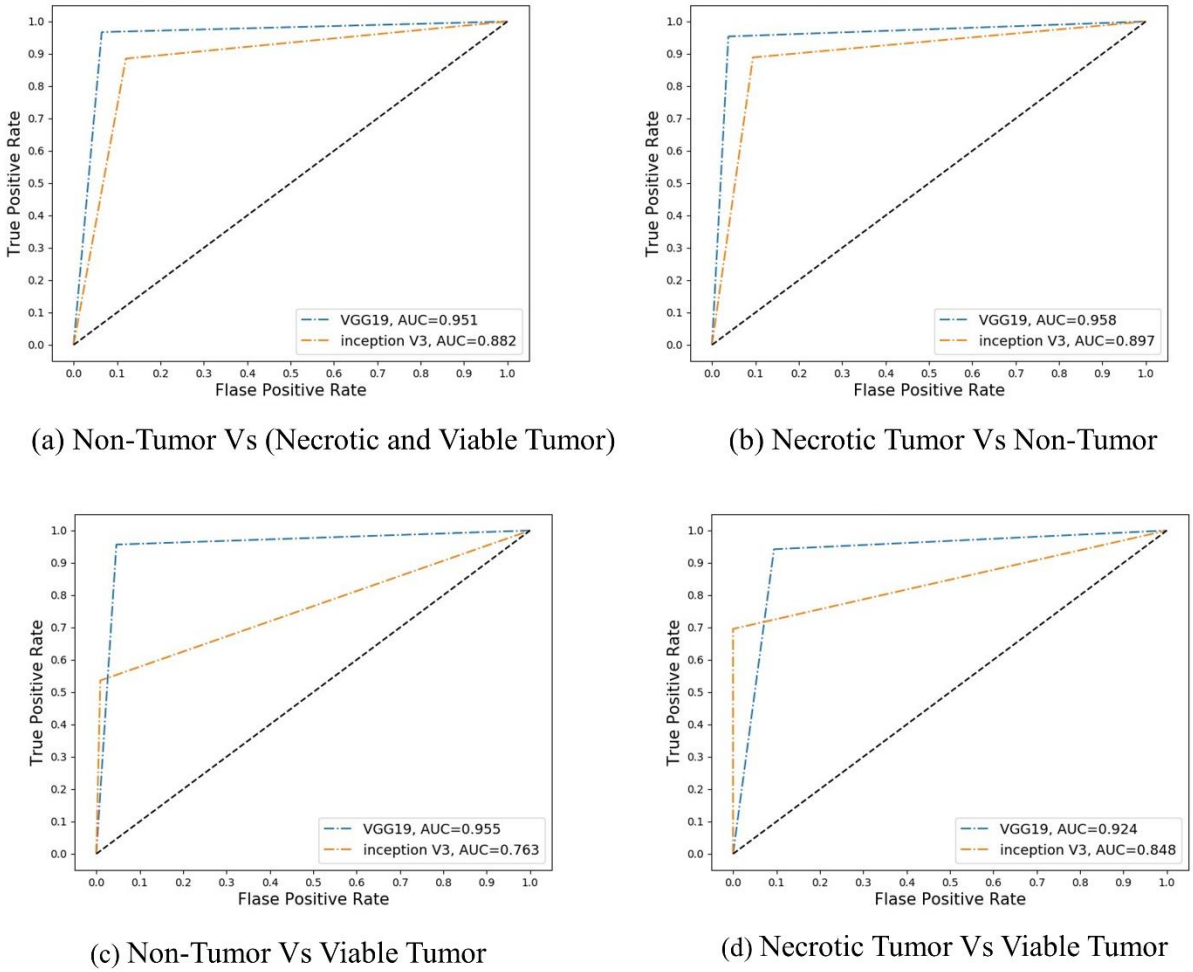


Figure 3.4: ROC and AUC of all two-class classifications on osteosarcoma classification.

To the best of our knowledge, this is the first pipeline that has been used in VGG19 and Inception architecture in Deep learning to recognize osteosarcoma malignancy. The adjusted model can identify the minimal differences in images to predict the early signs of cancer. If the pipeline was deployed in various medical facilities, our model could help pathologists as an adjunct tool by reducing their extensive work.

The VGG19 model achieves the best accuracy compared to Arunachalam et al.'s deep learning model (a CNN model with three pairs of convolutions and pulling layers for sub-sampling and two fully connected multi-layer perceptron). Table 3.4 represents the comparison of these two works. We have done a binary classification for all possible combinations between three classes, where Arunachalam et al. [51]'s deep learning model provides a direct class-specific accuracy. Therefore Table 3.4 represents our average accuracy for a specific tumor class. For example, for a viable tumor, the average of VT vs. NT and NCT vs. VT; for a necrotic tumor, the average of NCT vs. NT and NCT vs. VT; and for non-tumor, the average of NT vs. NCT and VT, NCT vs. NT, and VT vs. NT is represented. The comparison is made on the whole images (tile accuracy), as we have used the 1144 whole images for our classification. Table 3.4 shows a better performance of non-tumor than in other classes, which the imbalance data in each class may cause. This dataset contains 536, 345, and 263 whole images of non-tumor, viable, and necrotic tumors.

Table 3.4: Result comparison of osteosarcoma classification.

Tumor type	Tile accuracy in %	
	VGG 19	Arunachalam et al. [51]'s deep learning model
Non-Tumor	**95.45**	89.5
Necrotic Tumor	**94.34**	91.5
Viable Tumor	**94.26**	92.6

Limitations include the lack of evaluation from pathologists. Therefore, even though our model reaches a high performance, it is suggested that the tool should be used under a pathologist's supervision. A further study is to compare our model's performance with expert pathologists. The comparison can ensure this tool can detect new malignant cases in clinical practices. Besides, the existing data set might not indicate the future histological images from patients; therefore, the generalizability of our model might be problematic. To address this issue, it would be helpful to be adopted in medical facilities to assess its performance.

3.6 Conclusion

It is crucial to automate the classification of histological images by computer-aided systems within medical image processing. However, it is difficult and time-consuming to carry out the microscopic examination of histological images. Automatic histology diagnosis alleviates the workload and enables pathologists to focus on critical cases. In this work, we used two pre-trained networks from the Keras library, including VGG19 and InceptionV3. Regularization and optimization techniques were performed to avoid variance. The analyses were performed in two ways, one binary classification, and one multi-class classification. VGG19 model achieved the highest accuracy in both binary and multi-class classifications, with 95.65% and 93.91%, respectively. Furthermore, the highest F1 score in the binary class belonged to the Necrotic Tumor versus Non-Tumor, 0.97. Thus, our study has outperformed both binary and multi-class compared to the previous research on the same data. And finally, this study was the first usage of VGG19 and Inception V3 on the Osteosarcoma dataset, and the same framework can also be applied to other types of cancer.

Chapter 4

Wound Localization

4.1 Problem Statement

Wound localization begins with placing a bounding box around the wound or ulcer in a wound image and then cropping the bounded box for further processing. Figure 4.1 provides a brief overview of an automated wound analysis system. The extracted (cropped) region of a wound image will be passed as input to the segmentation and classification modules (named wound segmenter and wound classifier, respectively). The wound segmenter will segment a wound image for feature quantification (area, perimeter, width, height, etc.). The wound classifier will classify the image into different wound categories (DFU, PU, VLU, etc.). The classifier also classifies wound images into different tissue composition types (necrotic, slough, granulation, epithelium, and healed dermis) based on pixel colors. Wound localization can significantly simplify these subsequent wound analysis steps by removing unnecessary areas of wound images. Additionally, by limiting data capture to only the wounded tissue, all distinguishing features (face, tattoo, birthmark, etc.) are removed, enhancing patient privacy via wound localization. An example of extraneous data collection is demonstrated in Figure 4.1, in which the wound image also captures the bed sheet, pillow, calendar, fingers, etc. Non-wound image data can confuse the wound segmentation and classification algorithms resulting in reduced performance and weakened overall performance of the wound analysis intelligent system.

One of the most popular deep neural networks for image processing is the convolutional neural network (CNN). In CNN, the input is presented as a tensor in the shape of "number of images × width × height × depth." For a single image size of 1000 × 1000 pixels and three channels (RGB) in-depth, the input size to the CNN will be 3 million. If the first layer of the CNN has 1000 neurons, the requirement would be to train 1 billion parameters for this single layer CNN. Thus, it is prohibitively expensive to train these networks, given the limited hardware capacity. For this reason, we have to down-sample the input, which leads to loss of information and, ultimately, the network's poor performance. Using our wound localizer, we provide the segmenter and classifier with much smaller images, resulting in improved performance at significantly less cost.

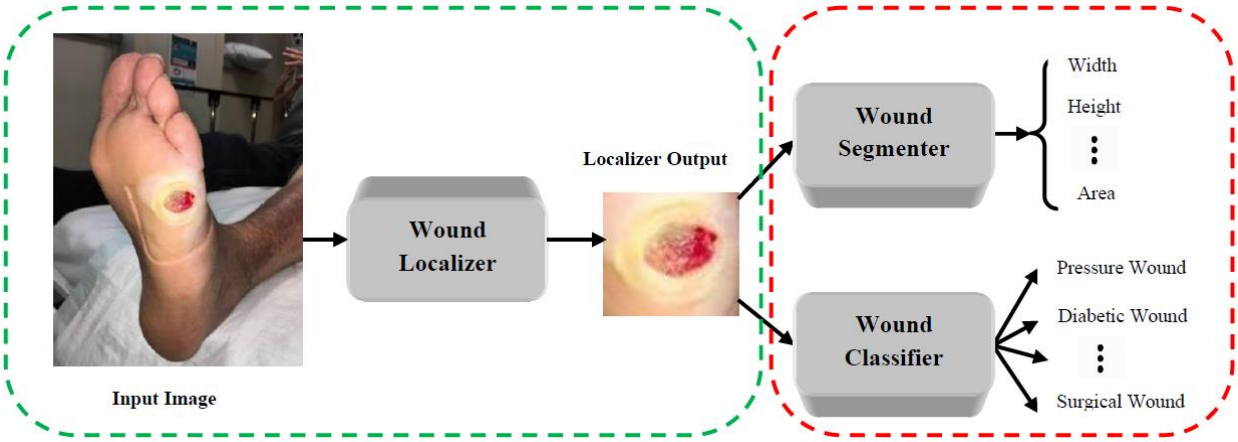


Figure 4.1: Automated wound system overview (the wound image is taken from the AZH wound dataset).

4.2 Related Works

Manually localized wound images are used as inputs to neural networks for wound segmentation. Papazoglou et al. [54], Hettiarachchi et al. [55], and Chang et al. [56] used manually localized wound images for their segmentation works. Wound and tissue classification also takes

manually localized wound images as inputs of neural networks. Wantanajittikul et al. [57], Goyal et al. [58], Shenoy et al. [59], Alzubaidi et al. [60], and Pinero et al. [61] used manually cropped ROIs for their wound type classification and wound tissue classification works.

Manual wound localization, as described above, is tedious and time-consuming, especially when a large amount of training data is considered in the subsequent tasks (segmentation, classification, etc.) using deep learning models. To the best of our knowledge, there are only two publications on automated wound localization with machine learning and deep learning models. Goyal et al. [62] proposed methods for Diabetic Foot Ulcer (DFU) detection and localization on mobile devices. They introduced a dataset including 1,775 DFU images and used SSD-MobileNet, SSD-InceptionV2, Faster R-CNN with InceptionV2, and R-FCN with Resnet 101 models for wound localization. They used mean Average Precision (mAP) and overlap percentage metrics to evaluate localization performance. From the mAP point of view, the best results were generated by Faster R-CNN with the InceptionV2 model, and from an overlap percentage point of view, the R-FCN with ResNet101 generates the best results. As shown in Section 5, there is still much space for improvement using the latest deep learning models. They used the Faster R-CNN with InceptionV2 model on an Android phone for smartphone applications. In another work, Goyal et al. [63] proposed a new dataset of DFUs and a classification method that predicts the presence of infection or ischemia in the DFU. For these experiments, they introduced a dataset including 1,459 DFU images. Their data augmentation step used Faster-RCNN and InceptionResNetV2 architectures for ROI detection. However, no evaluation metric is presented for wound ROI detection on their dataset.

4.3 Methodology

4.3.1 Data Collection

The wound dataset has been collected from the AZH Wound and Vascular Center, Milwaukee, WI, USA. This dataset (AZH Wound Database) contains 1,010 wound images. Three types of ulcers have been included in the dataset: Diabetic foot ulcer (DFU), Pressure Ulcer (PU), and Venous Ulcer (VU). All the images are captured with iPad and DSLR cameras. No specific environmental or illumination condition has been applied during image capturing. These images are further processed and used as training and test data. The dataset is available at https://github.com/uwm-bigdata/wound_localization. Additionally, for testing the robustness and reliability of our models, 52 images have been downloaded from Medetec Wound Database [64]. Though this database contains all types of open wound images such as abdominal wounds, burns and scalds, diabetic foot ulcers, haemangiomas, venous ulcers, arterial ulcers, malignant wounds, and more; we have only collected three types: diabetic foot ulcer, a venous ulcer, and pressure ulcer images for training. Finally, we collected more data from the AZH wound center and Metedec Wound Database to create a bigger and mixed database (BMAZHM Wound Database). The BMAZHM dataset contains 1010 images from the AZH wound center and the newly collected 790 images, where 538 are collected from the AZH wound center, and 252 are collected from the Metedec database.

4.3.2 Data Preparation

Our models can take different width-height ratios of images in both training and test datasets, so we do not make the image size uniform. To increase the number of images for our dataset, we have applied augmentations on the AZH Wound Database with rotation, flipping, and blurring, resulting in 4,050 image data. The rotation augmentations include rotations of each image for ±25 degrees, and the flipping augmentations include flipping each image horizontally and vertically. Some specific augmentations (e.g., blurring) are not applied to every image. They might have lost some features, which leads to an augmented dataset without an exact multiplier of the original images. The training and testing datasets were separated before the augmentation to ensure no overlapping occurred. For the training dataset, we have 3,645 images, and for the testing dataset, we have 405 images. All the collected images have been labeled manually for the training and evaluation of our models. We have used an MIT-licensed free graphical image annotation tool named labelImg [65] for data labeling. Annotations are saved in YOLO format as a text file for each image, containing the class number, center coordinates of the bounding box(s), and height and width of the bounding box(s). Annotations are converted to Pascal VOC format and, together with images, passed as the inputs to the SSD model. The following augmentations are applied on newly collected 790 images of the BMAZHM Wound Database: rotation in ±25 degrees, horizontal and vertical flipping, blurring, and addition of Gaussian noise. After augmentation, there are 5530 wound images. Previously augmented data of 4050 images are mixed with it, which gives us a total of 9580 images. These images are split into 7664, 1404, and 512 for training, validation, and testing. The split is done carefully without overlapping train, validation, and test datasets. The data labeling of newly collected and augmented images is performed using the same tool labelImg [65], and the annotations are saved in YOLO format.

4.3.3 Model Training

We used YOLOv3 and SSD as our wound localization models. These models are selected for their popularity, reliability, and time management for object detection. A comparison of these two models for our wound detection task has been presented in the result and discussion section.

We used a single class named "wound" for our model training. For the YOLOv3 model, we used the YOLO annotations. The model is trained for 273 epochs with a batch size of 8, a learning rate of 0.001, and stochastic gradient descent (SGD) optimizer. The YOLOv3-416 model is used for wound bounding box detection. In SSD, we used the Pascal VOC annotations converted from YOLO annotations for the model training. The SSD model is trained for 475 epochs with a batch size of 8, a learning rate of 0.001, and stochastic gradient descent (SGD) optimizer. The SSD300 model is used for wound bounding box detection. During the detection process, the image size is set to 416, and the IoU threshold is set to 0.5. The models are written in Python programming language using the Pytorch deep learning framework and trained on an Nvidia GeForce RTX 2080Ti GPU platform. The trained YOLOv3 model is retrained using the BMAZHM dataset. The model is trained for 273 epochs with a batch size of 32, a learning rate of 0.01, and stochastic gradient descent (SGD) optimizer. Google Colab platform with a Tesla V100-SXM2-16GB GPU is used for training.

4.4 Result and Discussion

4.4.1 Performance Metrics

We have adopted Precision (equation 2.2), Recall (equation 2.3), F1-score (equation 2.4), Intersection over Union (IoU) (equation 2.1), and the Mean Average Precision (mAP) (equations 2.5 and 2.6) as the evaluation metrics to evaluate the localization performance. To define the Precision, Recall, and F1-score, we set the IoU threshold to 0.5. If IoU > 0.5 and the wound is correctly classified, the result is true positive; correctly classified means both the predicted label and the actual label are both wounds. If IoU > 0.5 and the wound is wrongly classified, the result is false negative; wrongly classified means the predicted label is not a wound, but the actual label is a wound. If IoU < 0.5 and the wound is correctly classified, the result is true negative; correctly classified means the predicted label and the actual label are non-wounds (i.e., skin, background, etc.). If IoU < 0.5 and the wound is wrongly classified, then the result is false positive; where wrongly classified means the predicted label is a wound, but the actual label is not a wound. Precision and Recall show the accuracy of our localization model.

4.4.2 Result

We tested the performance of our models with the test dataset of 405 wound images from the AZH Wound Database. For the YOLOv3 model, with an IoU of 0.5 and non-maximum suppression of 0.45, we get the mAP value of 0.939. The Precision, Recall, and F1-score for the YOLOv3 model are 0.925, 0.905, and 0.915, respectively. For the SSD model, with an IoU of 0.5 and non-maximum suppression of 0.45, we get the mAP value of 0.864. The SSD model's

Precision, Recall, and F1-score are 0.902, 0.584, and 0.709, respectively. For the BMAZHM dataset, YOLOv3 achieves an mAP value of 0.973, where the IoU is 0.5. The Precision, Recall, and F1-score for the YOLOv3 model are 0.939, 0.960, and 0.949, respectively. Table 4.1 shows a summary of our evaluation results.

Table 4.1: Result summary for wound localization.

Database	Network	Precision	Recall	F1-Score	mAP
AHZ Wound Database	YOLOv3	0.925	0.905	0.915	0.939
	SSD	0.902	0.584	0.709	0.864
BMAZHM Wound Database	YOLOv3	0.939	0.960	0.949	0.973

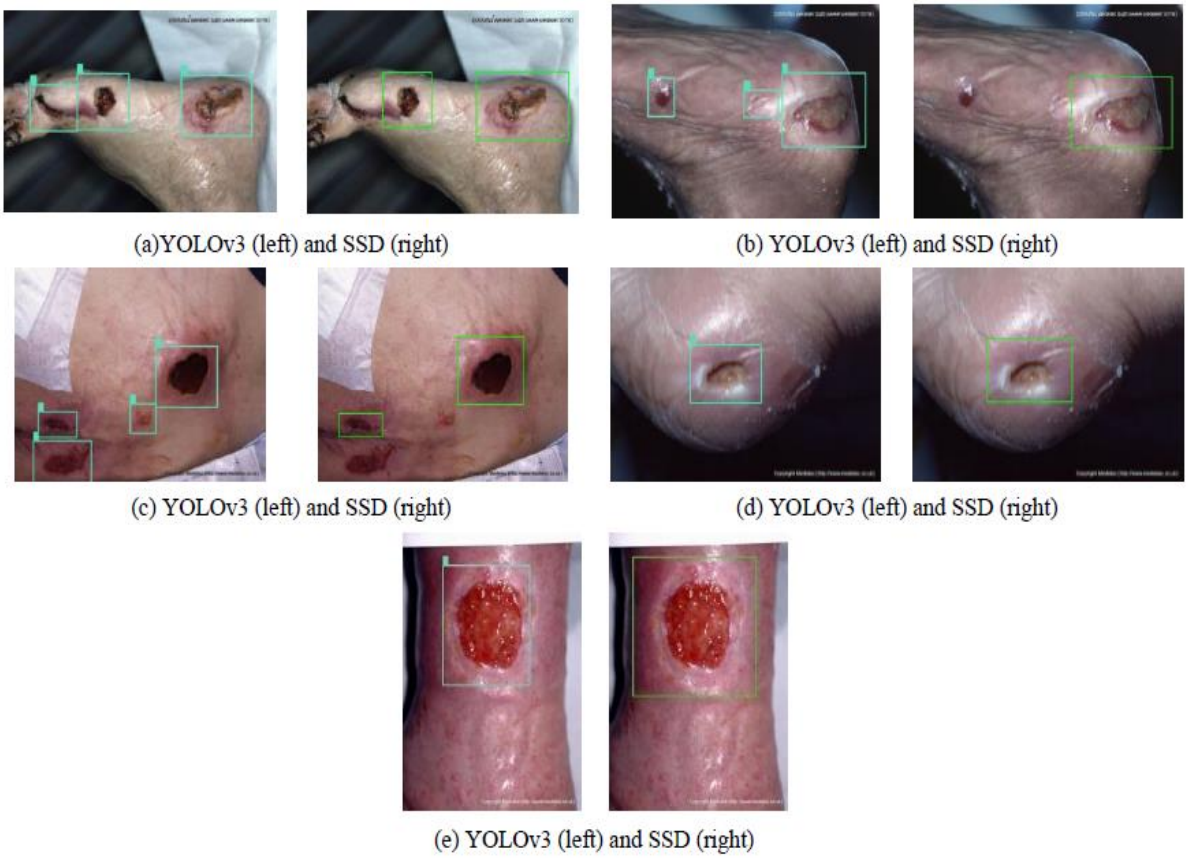


Figure 4.2: Robustness and reliability testing output of the wound localizer.

The robustness and reliability testing on the Medetec dataset shows a promising result. With our best model (YOLOv3 according to our AZH Wound Database evaluation), the Precision, Recall, F1-score, and mAP values are 0.926, 0.603, 0.73, and 0.808, respectively. Some of the testing outputs with the YOLOv3 and SSD models are shown in Figure 4.2.

4.4.3 Discussion

From the results shown above, YOLOv3 gives significantly better results. The mAP value for the YOLOv3 model is higher than the SSD model, with a difference of 7.5% on the AZH dataset. All the evaluation metrics (Precision, Recall, and F1-score) reflect better values for the YOLOv3 model than for the SSD model. From Table 4.1, SSD reflects a low Recall and high Precision, which leads to the decision that SSD is a very picky or fault-finding model. Most images detected as wounds are actual wounds, but it also misses a lot of actual wounds. On the other hand, Table 4.1 shows that the YOLOv3 model has a high precision (0.925) and high recall (0.905) value, representing a better and more stable model. The performance metrics are improved significantly for the YOLOv3 model using a more extensive dataset (BMAZHM Wound Database).

Our YOLOv3 model achieves a promising result with an mAP value of 0.808 regarding the robustness and reliability test. This evaluation result is reasonable, with our model trained on AZH Wound Database and Medetec being a completely new and unseen dataset. From Figure 4.2, we can see that the YOLOv3 model produces better results than the SSD model. In Figure 4.2, from (a), (b), and (c), we can see that the SSD model misses some wound ROIs in the case of an image with multiple wounds, which shows the picky behavior of the SSD model as discussed

above. From 5.2(d) and 5.2(e), we can see that both models perform well, but the SSD model captures slightly more healthy skins than the YOLOv3 model. So, we can confidently say that the YOLOv3 model does better wound localization than the SDD model.

Both the YOLOv3 model and the Tiny-YOLOv3 model perform better than Goyal et al.'s wound localization work [62], where they achieved an mAP value of 0.849, 0.872, 0.918, and 0.906 for SSD-MobileNet, SSD-InceptionV2, Faster R-CNN with InceptionV2, and R-FCN with Resnet 101 models, respectively. Their dataset consists of only diabetic foot ulcer images (1775), but our AZH dataset contains three types of ulcer images (1010), and this comparison may vary depending on the dataset. As the dataset of [62] is not publicly accessible, we implemented their best model (Faster R-CNN with InceptionV2) on our AZH Wound Database. This model uses Inception V2 for feature extraction and Faster R-CNN for object localization. Figure 4.3 shows the comparison of the Faster R-CNN with the InceptionV2 model with our YOLOv3 model and SSD model. The YOLOv3 model performs better than Goyal et al.'s best model on the AZH dataset. In general, Darknet-53 (used by YOLOv3 for feature extraction) is much newer and better [66] than InceptionV2 [67], and the same claim goes for YOLO layers against Faster RCNN layers [68], which reflects in Figure 4.3. The YOLOv3 model is improved further by training with a more significant and improved BMAZHM Wound Database, which achieved an mAP of 0.97 and outperformed Goyal et al.'s wound localization work by a considerable margin.

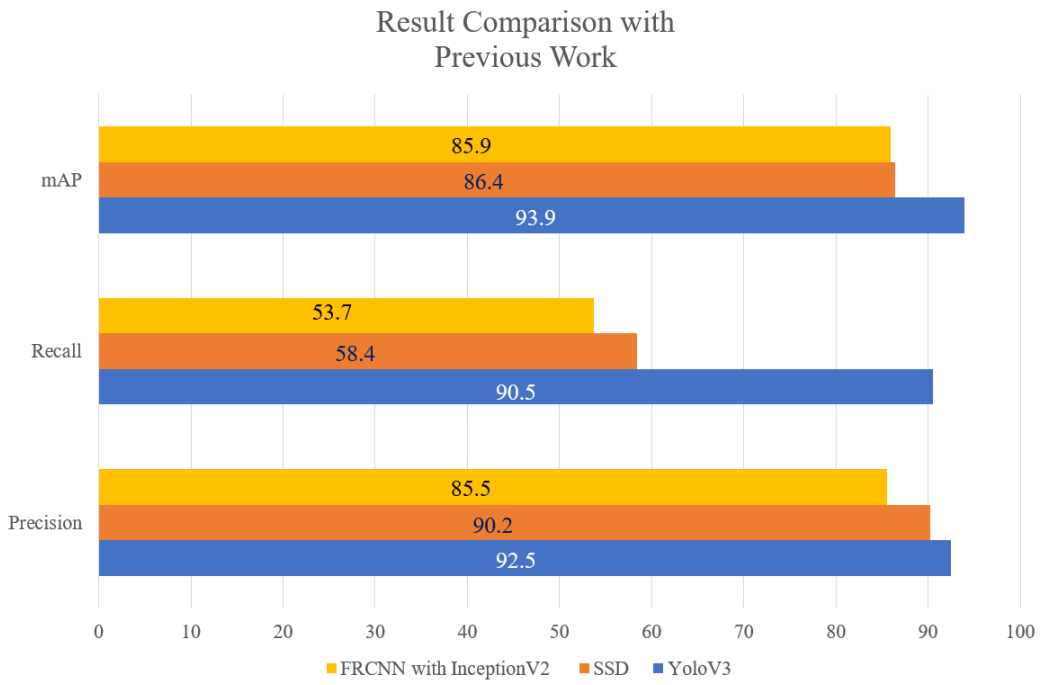


Figure 4.3: The Faster R-CNN with InceptionV2 (the best model investigated in [62]) is compared with two models (YOLOv3 and SSD) employed in our work. All three models are trained and tested on the same dataset, the AZH Wound Database.

Though there may be some downsides to applying localization as the first step: errors made by the localization system may propagate to the final prediction (compared to an end-to-end system) and the context around the wound (e.g., wound location, healthy skin information, and more.) can get lost; the ROI localization can significantly make the classifier and segmenter easier, improve their performances, and ensure data privacy. Thus, the proposed ROI detection as the first step in wound analysis provides enormous benefits despite its potential downsides.

4.5 Conclusion

This chapter focuses on building an automated wound localizer, the first step of creating an intelligent wound diagnostic system. The outcome of the localizer will be the input of subsequent wound processing tasks, such as wound segmentation and classification. Our system initially achieved the mAP value of 0.939 on the AZH dataset and a significantly higher mAP value of 0.973 on a more prominent and improved BMAZHM wound dataset and outperforms the only existing automated wound localization work [62] by a considerable margin. Furthermore, we have automated our wound localizer, which is unique compared to most previous works based on localizing wounds manually from the original image. The present system has great importance in future research on intelligent wound healing. Although we achieve very high mAP values, the developed model may completely miss some wound ROIs (i.e., false negatives). Fortunately, our model does not detect partial ROIs, meaning that a bounding box always covers the whole wound if found, which always provides entire wound regions for the subsequence tasks.

Chapter 5

Wound Image Classification

5.1 Burn Wound Classification

5.1.1 Problem Statement

Management of acute and chronic wounds is a challenge and burden to healthcare systems. A recent retrospective analysis of Medicare beneficiaries in the United States demonstrated that ~8.2 million people had acute and chronic wounds at an annual cost ranging from $28.1 billion to $96.8 billion [18]. One of the significant wound types is burn wounds. This section demonstrates two types of burn wound classifications: 1) binary classification between the graft (full-thickness + deep dermal) and non-graft (superficial dermal) burns and 2) multi-class classification among full-thickness, deep dermal, and superficial dermal burn depths. Deep learning methods are used to perform these classifications.

5.1.2 Related Works

Several machine learning-based classifiers are used for burn image classifications. An SVM-based machine learning method has been developed by Yadav et al. [69] for burn diagnosis from burn images in the BIP_US database. They used color, texture, and depth features to train their classifier to classify three types of burns: superficial dermal, deep dermal, and full-thickness burn. Abubakar et al. [70] proposed a machine learning-based approach to distinguish between burn wounds and pressure ulcers. They used pre-trained deep architectures like VGG-face, ResNet101, and ResNet152 to extract the features from the images and then fed them into an SVM

classifier for classifying the images into burn or pressure wound classes. The dataset used in this study includes 29 pressure and 31 burn wound images obtained from the internet and a hospital. After augmentation, they had three categories: burn, pressure, and healthy skin, with 990 sample images in each class. Several experiments, including binary classification (burn or pressure) and 3-class classification (burn, pressure, and healthy skin), were conducted with the accuracy value of 99.9% as their best classification accuracy using ResNet152 for both binary and 3-class classification problems.

Three types of burn depths (superficial dermal, deep dermal, and full-thickness) were detected by Serrano et al. [71]. They used their own dataset. First, they segmented the images by transforming them into the L*u*v* color space and applying filter smoothing, distance calculation, and thresholding to the new images. Next, they extracted 16 color and texture descriptors from the segmented images and fed them to the fuzzy-ARTMAP neural network for three-type burn classification. In another study, Serrano et al. [72] used Multidimensional Scaling Analysis (MDS) to obtain physical features from images, which are then translated into mathematical features such as chroma, outliers, hue, skewness, and kurtosis. These features are then passed to the SVM classifier to classify burns that do not need grafts (heal spontaneously) and burns that need grafts. Burns_BIP_US Database was used for their experiment, and they used 20 images for training and 74 images for testing. Though they achieved high sensitivity (0.97), the accuracy (79.73%) and specificity (0.60) were poor. Acha et al. [61] classified burn images into five classes: superficial dermal (blisters), superficial dermal (red), deep dermal, full-thickness (beige), and full-thickness (brown); by using color and texture features. They used their own developed dataset, where 62 images were collected with a digital photographic camera. Using a sequential backward selection (SBS) method, they selected six features (lightness, hue, SD of hue, u* chrominance, SD of v*,

and skewness of lightness). Then, they fed them to a Fuzzy-ARTMAP neural network for five-class wound classification. Acha et al. [73] used MDS to obtain physical features (amount of pink, texture of the color, and colorfulness) from 2D images, which were then translated to mathematical features and passed to the KNN classifier to classify three types of burn (superficial dermal, deep dermal, and full-thickness). They used the Burns_BIP_US Database to perform binary and multiclass classifications, where they used 20 images for training and 74 images for testing. They achieved an 83.8% accuracy while classifying between graft and non-graft classes and a 66.2% accuracy while classifying among the three classes (superficial dermal, deep dermal, and full-thickness).

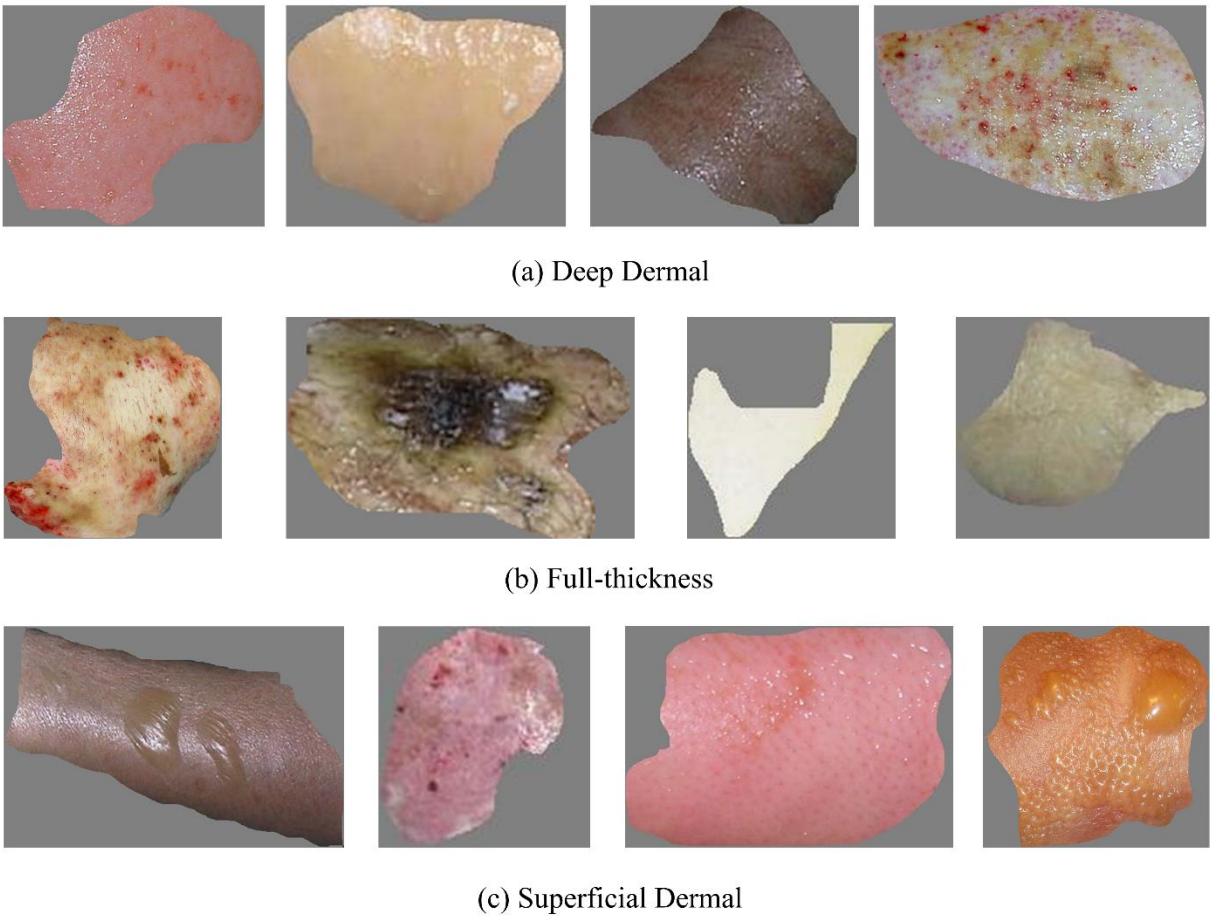
Deep learning was used by Despo et al. [74] for efficient burn classification, where they built an end-to-end deep learning model independent of any specific features of burn wounds. Also, they proposed a new wound dataset including 929 images, most of which were obtained from a medical center, while the rest were collected from the internet. They defined four labels for the dataset images: superficial, superficial/deep partial thickness, full-thickness, and undebrided. The authors classified the images into burn or no burn categories in the first step. The second step was related to separating the wound area from the rest of the image. In the final step, they classified the burn wound into different classes showing the depth levels of each one. The authors used a modified fully convolutional network (FCN) built upon the VGG-16 network.

5.1.3 Methodology

5.1.3.1 Dataset

This experiment uses a public database named Burns_BIP_US database [75]. This dataset contains 94 images from three burn wound types: full-thickness, deep dermal, and superficial

dermal. There are 20, 32, and 42 samples in each class, respectively. The images have jpg and bmp formats, and they are in different sizes. Figure 5.1 shows some dataset samples.



(a) Deep Dermal

(b) Full-thickness

(c) Superficial Dermal

Figure 5.1: BIP US database sample images. (a) represents deep dermal images, (b) represents full-thickness images, and (c) represents superficial dermal images.

5.1.3.2 Data Preparation

In the 3-class classification problem, we split the data into train, validation, and test sets with 76, 9, and 9 images. The splitting is done randomly. For the binary classification experiment, we followed the same data splitting strategy of [69], [72], and [73], as we want to compare our method with these approaches. Thus, we put 74 images into the test set and the rest of the data samples into the train and validation sets. As a result, the binary classification contains graft and

non-graft, where graft contains full-thickness and deep dermal images, and non-graft contains superficial dermal images.

The training data were augmented by generating 15 images from each image using transformations like rotating, flipping, cropping, and mirroring. Therefore, for the 3-class classification experiment, the number of training samples in the classes deep dermal, full-thickness, and superficial dermal increased to 416, 224, and 576 after augmentation. Thus, we ended up with 128 images in the non-grafted class and 144 in the grafted category for the binary classification case. Table 5.1 shows the dataset splitting and augmentation statistics.

Table 5.1: Database summary for burn classification.

		Class	**Train**	**Validation**	**Test**	**Total**
Binary Classification Dataset	**Original**	Graft	9	2	41	**52**
		Non-Graft	8	1	33	**42**
		Total	**17**	**3**	**74**	**94**
	Augmented	Graft	144	2	41	**187**
		Non-Graft	128	1	33	**162**
		Total	**272**	**3**	**74**	**349**
Multiclass Classification Dataset	**Original**	Full-thickness	14	3	3	**20**
		Deep dermal	26	3	3	**32**
		Superficial dermal	36	3	3	**42**
		Total	**76**	**9**	**9**	**94**
	Augmented	Full-thickness	224	3	3	**230**
		Deep dermal	416	3	3	**422**
		Superficial dermal	576	3	3	**582**
		Total	**1216**	**9**	**9**	**1234**

5.1.3.3 Models

Some popular transfer learning and end-to-end learning models are used to classify this dataset. Transfer learning models include AlexNet [28], VGG16 [29], VGG19 [29], InceptionV3 [30], and NasNet Large [35] networks. The top layers are removed in all transfer learning networks, and trained layers are frozen. Some Dense layers and one output layer are added on top of these networks. End-to-end learning models are Alex Net, SimpleNet, and DeepNet, where the Alex Net is trained from scratch. The SimpleNet and DeepNet are two own developed networks where the number of Convolution and Max-pooling layers are small and large in numbers, respectively. Figure 5.2 and Figure 5.3 show the architecture of SimpleNet and DeepNet networks, respectively. SimpleNet contains four convolution and max-pooling layers with a different kernel and depth sizes and two fully connected layers. DeepNet contains five convolution and max-pooling layers with a different kernel and depth sizes and three fully connected layers. All convolution and fully connected layers have the ReLU activation.

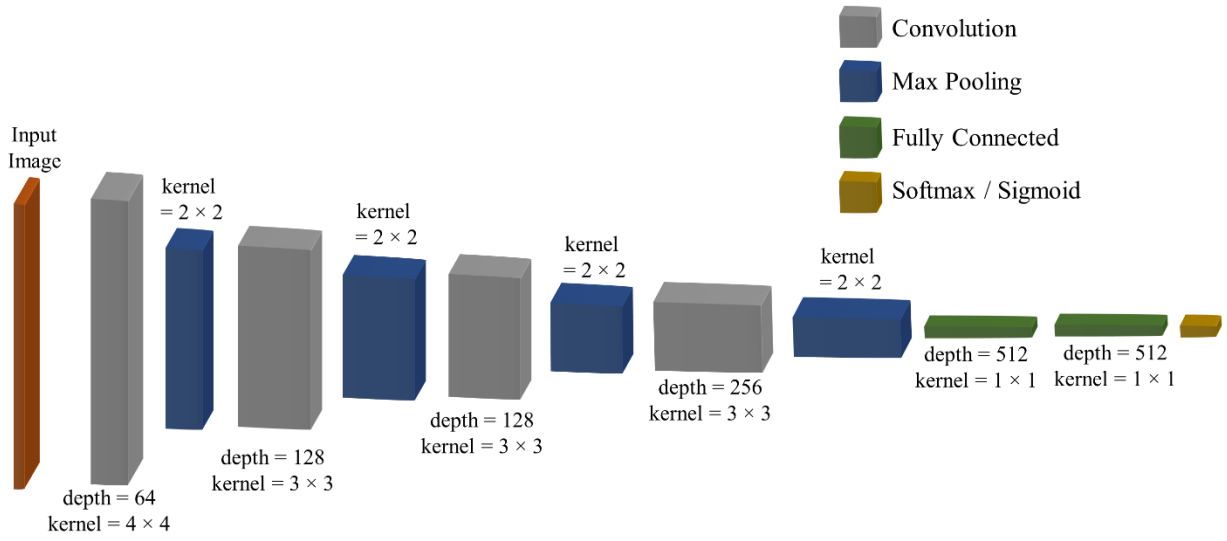


Figure 5.2: SimpleNet architecture.

Except for the transfer learning of the ALexNet network, all models are written in Python programming language using Tensorflow.Keras deep learning framework and trained on an Nvidia GeForce RTX 2080Ti GPU platform. Transfer learning of AlexNet was implemented in version R2020a of MATLAB software and trained on an NVIDIA GEFORCE MX 150 GPU with 2GB of memory.

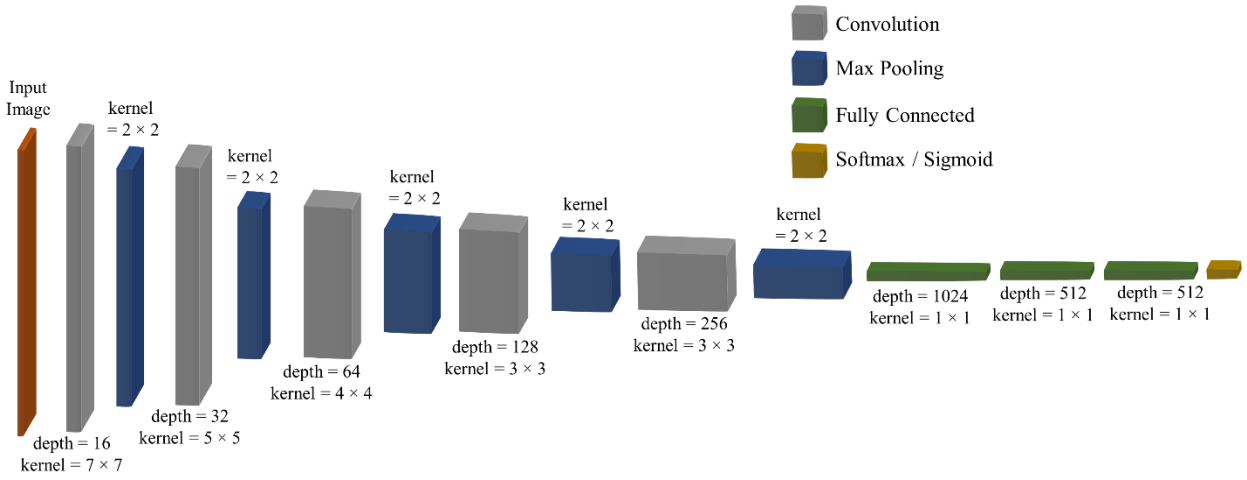


Figure 5.3: DeepNet architecture.

5.1.4 Results and Discussion

5.1.4.1 Result

We use accuracy (equation 2.7) as the performance metric to investigate the classification performance. In addition, precision (equation 2.2), recall (equation 2.3), f1-score (equation 2.4), and AUC values are also reported for the binary classification. For multiclass classification, we also presented a confusion matrix of best models. The binary and multiclass classifications results are shown in Tables 5.2 and 5.3.

Table 5.2: Burn image binary classification results.

Model Type	Model	Accuracy	Precision	Recall	F1-Score	AUC
Transfer Learning	Alex Net	90.50%	90.6%	87.9%	89.22%	0.913
	VGG16	75.68%	70.27%	78.79%	74.29%	0.760
	VGG19	70.27%	66.67%	66.67%	66.67%	0.699
	InceptionV3	71.62%	73.08%	57.58%	64.41%	0.703
	NasNet Large	64.86%	57.44%	81.82%	67.50%	0.665
End-to-End Learning	Alex Net	77.03%	76.67%	69.70%	73.02%	0.763
	SimpleNet	87.84%	85.29%	87.88%	86.57%	0.878
	DeepNet	83.78%	88.89%	72.73%	80.00%	0.827

Table 5.3: Burn image multiclass classification results.

Model Type	Model	Accuracy
Transfer Learning	Alex Net	77.78%
	VGG16	44.44%
	VGG19	33.33%
	InceptionV3	22.22%
	NasNet Large	33.33%
End-to-End Learning	Alex Net	77.78%
	SimpleNet	44.44%
	DeepNet	66.67%

5.1.4.2 Discussion

From Table 5.2, the AlexNet with transfer learning model performs the best with 90.5% accuracy. On the other hand, our developed end-to-end learning achieves the second-best accuracy with the SimpleNet network. The same result pattern can be seen in precision, recall, f1-score, and AUC scores, where AlexNet and SimpleNet hold the best and second-best scores. One reason for not achieving more accuracy is the scarcity of data. We only have 17 training data and three validation data, which is very few for deep learning models.

From Table 5.3, for multiclass burn image classification, the AlexNet with both transfer learning and end-to-end learning performs the best with 77.78% accuracy. DeepNet networks achieve the second-best accuracy of 66.67%. Again, the reason behind this poor performance is the scarcity of data. We only have 9 validation data, and the model is tested on 9 images which are very few for deep learning models.

This result shown in Table 5.2 and Table 5.3 indicates that lightweight networks performed best in both binary and multiclass classifications. On the other hand, large networks perform the worst. There is a smooth curve of decreasing accuracy (except for two cases) from light to deep neural networks. Figure 5.4 shows this behavior of deep learning models. Here, AlexNet has the highest accuracy for both binary and multiclass classifications, which contains only 7 layers, including 3 convolution layers, 2 max-pooling layers, and 2 fully connected layers. SimpleNet has the second-highest accuracy for binary classification and the third-highest accuracy for multiclass classifications, which contains only 10 layers, including 4 convolution layers, 4 max-pooling layers, and 2 fully connected layers. Followed by DeepNet, which includes 13 layers, including 5 convolution layers, 5 max-pooling layers, and 3 fully connected layers. The decreasing accuracy is followed by VGG16, VGG19, NasNetLarge, and InceptionV3 networks for both classifications

(the only exception in InceptionV3 in binary classification), which are 16, 19, 22, and 48 layers deep, respectively.

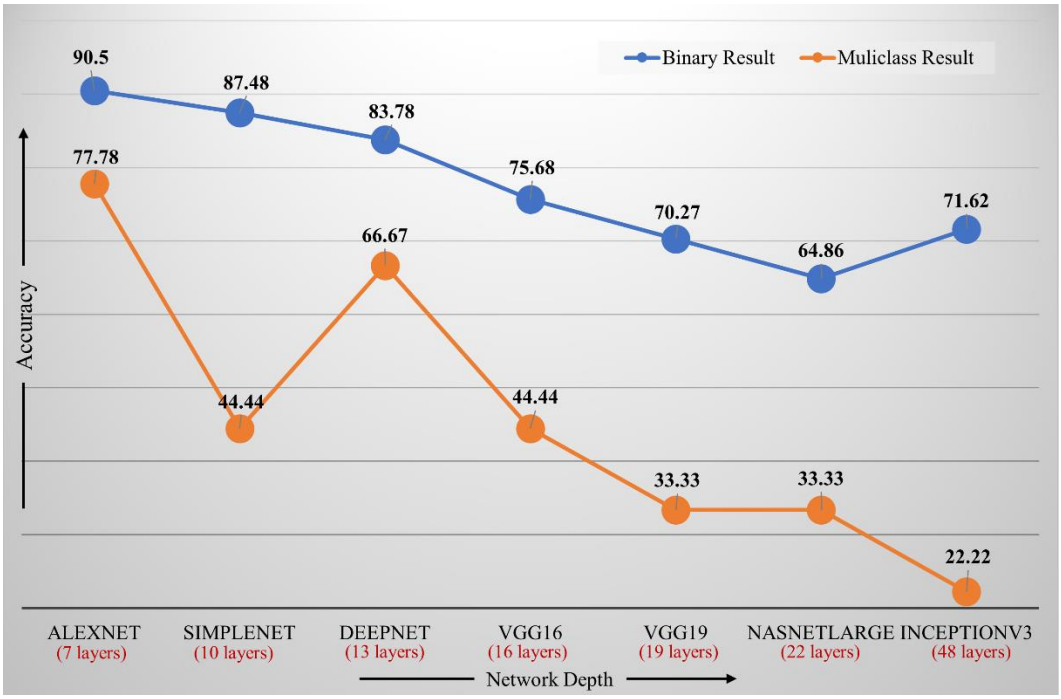


Figure 5.4: The accuracy versus network depth comparison for burn classification.

The confusion matrix of the top three models (AlexNet with transfer learning, AlexNet with end-to-end learning, and DeepNet) is shown in Figure 5.5. AlexNet with transfer learning classifies all the deep dermal images correctly, but it misclassifies one full-thickness image as superficial dermal and one superficial dermal image as deep dermal. AlexNet, with end-to-end learning, also classifies all the deep dermal images correctly. Still, it misclassifies one full-thickness image as deep dermal and one superficial dermal image as deep dermal. Finally, DeepNet misclassifies one deep dermal image as superficial dermal, one full-thickness image as deep dermal, and one superficial dermal image as deep dermal. Some of the misclassifications by these networks are shown in Figure 5.6.

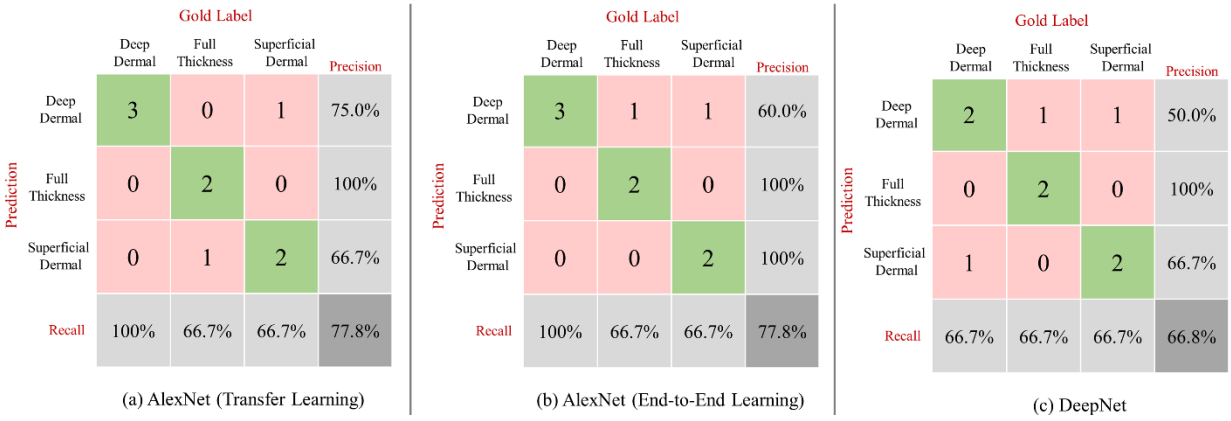
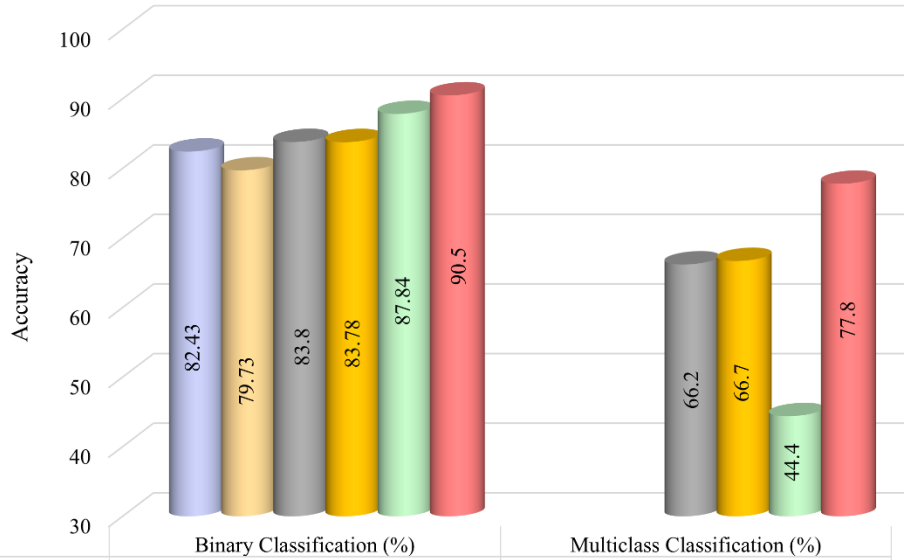


Figure 5.5: Confusion Matrix of best models for burn wound classification.

Figure 5.6: Examples of burn wound image misclassifications by the best models.

We followed the work of [69], [72], and [73] for the dataset preparation of the binary classification. In [69], the authors acquired an accuracy value of 82.43 %, as well as precision, recall, and F1-score values of 0.82, 0.88, and 0.85, respectively, using a traditional color-based method for feature extraction and support vector machine (SVM) for classification. In [72] authors achieved an accuracy of 79.73% using Multidimensional Scaling Analysis (MDS) method for feature extraction and SVM for classification. In [73] authors achieved an accuracy of 83.8% using

MDS method for feature extraction and KNN for classification. They also achieved an accuracy of 66.2% on multiclass classification using the same approach. For binary classification, AlexNet, SimpleNet, and DeepNet outperformed the existing works by a good margin. The AlexNet and DeepNet outperformed the only existing multiclass classification work [73] performed on this Burns_BIP_US database. A comparison of the existing work with our best models is shown in Figure 5.7. We claim that our top classifiers produced higher performances for the binary and multiclass classification tasks when all the supplied metrics were considered.



	Binary Classification (%)	Multiclass Classification (%)
Color_Feature+SVM [69]	82.43	
MDS+SVM [72]	79.73	
MDS+KNN [73]	83.8	66.2
DeepNet	83.78	66.7
SimpleNet	87.84	44.4
AlexNet	90.5	77.8

Figure 5.7: Performance comparison with existing works for burn wound classifications.

The performance of deep learning models is closely related to the amount of data. The Burns_BIP_US database contains only 94 images, which is really low for training a deep learning model. We only used 17 original images for training and 3 images for validation for binary classification. When training the model, we reached 100% train and 100% validation accuracy

very quickly (within 10 epochs), saving a model with 100% validation accuracy but learning very few features from the data, which leads to poor test accuracy. So, we applied a specifically designed callback that saves the model with 100% validation and train accuracy and considers the lowest validation loss during saving. The lowest validation loss we got is 4.26×10^{-14} but it still does not help with the test accuracy. For multiclass classification, we have only 76 original images for training and 9 images for both validation and test. After augmentation, we have 272 images for binary classification and 1216 images for multiclass classification, with 15 augmentations applied to each original image. But augmentation does not produce a new image; it just modifies the image properties to a specific extent with rotation, mirroring, etc. With this augmentation, we can not change (should not change as it will ruin the original class properties) some specific properties of the original image like color, texture, etc. So, to improve the classification performance, we need to increase the number of original images in the dataset. One future direction to increase the number of images is to apply Generative Adversarial Networks (GANs) [76] to generate synthetic data (images). But this is also highly dependent on the GAN model's performance and the acceptance of these images by the medical professionals (physicians). We have already applied the GAN model to generate synthetic electronic health record (EHR) data to perform wound healing prediction, improving model performance [77]. So, applying GANs to generate artificial burn wound images is a very good future option to improve the classification performance of this project.

5.2 Wound Severity Classification

5.2.1 Problem Statement

Wound severity classification is an essential part of the wound diagnosis process as this study can help physicians make quick and proper decisions on treatment plan making. There are three classes named green, yellow, and red. Green represents the wounds in the primary stage and is most likely to heal with proper treatment. The yellow class contains those wounds that need more attention and treatment than those in the green category. Finally, the red class contains the most severe wounds that require immediate action and quick and proper treatment. The characteristics upon which wound severity can be determined are shown in Figure 5.8.

5.2.2 Related Works

Some machine learning-based methods are applied for wound type classification. Goyal et al. [63] used traditional machine learning, deep learning, and ensemble CNN models for binary classification of ischemia versus non-ischemia and infection versus non-infection on DFU images. The authors developed a dataset containing 1459 DFU images that two healthcare professionals labeled. The authors used BayesNet, Random Forest, and Multilayer perceptron for traditional machine learning. Three CNN networks were used as deep-learning approaches. The ensemble CNN contains an SVM classifier that takes three CNN networks' bottleneck features as input. The test evaluation shows that traditional machine learning methods perform the worst, followed by deep-learning networks, while the ensemble CNN performs the best in both binary classifications. Abubakar et al. [70] proposed a machine learning approach to differentiating burn wounds and

pressure ulcers. The dataset used in this study includes 29 pressure and 31 burn wound images obtained from the internet and a hospital, respectively. Features were extracted using pre-trained deep architectures like VGG-face, ResNet101, and ResNet152 from the images and then fed into an SVM classifier to classify the images into burn or pressure wound classes. Several experiments were conducted, including binary classification (burn or pressure) and 3-class classification (burn, pressure, and healthy skin).

Characteristic	Red	Yellow	Green
Color			
Red 100%			X
Yellow-Grey <50%		X	
Yellow-Grey 50-100%	X		
Black-Brown	X		
Peri-wound			
Normal			X
Callus		X	
Red <1cm		X	
Red >1cm	X		
Maceration		X	
Maceration and Breakdown	X		
Size			
2cm or less			X
2 – 5 cm		X	
>5 cm	X		
Depth			
Minimal-None			X
1cm or less		X	
>1cm	X		

Figure 5.8: Photo characteristics of Red-Yellow-Green stratification.

Deep learning-based methods are also applied for wound type classifications. A novel CNN architecture named DFUNet was developed by Goyal et al. [58] for binary classification of healthy skin and DFU skin. A dataset of 397 wound images was presented, and data augmentation techniques were applied to increase the number of images. The proposed DFUNet utilized concatenating the outputs of three parallel convolutional layers with different filter sizes. Shenoy et al. [59] proposed a CNN-based method for binary classification of wound images. This study used a dataset of 1335 wound images collected via smartphones and the internet. The authors considered nine different labels (wound, infection (SSI), granulation tissue, fibrinous exudates, open wound, drainage, steri strips, staples, and sutures) for the dataset, where for each label, two subcategories (positive and negative) were considered. The authors used a modified VGG16 network named WoundNet as the classifier, pre-trained using the ImageNet dataset. In addition, the researchers created another network called Deepwound, an ensemble model that averages the results of three individual models. A binary patch classification of normal skin versus abnormal skin (DFU) is performed by Alzubaidi et al. [60] with a novel deep convolutional neural network named DFU_QUTNet. First, the authors introduced a new dataset consisting of 754-foot images from a diabetic hospital center in Iraq. The proposed network is a deep architecture with 58 layers, including 17 convolutional layers. The performance of their proposed method was compared with those of other deep CNNs like GoogLeNet, VGG16, and AlexNet.

A CNN-based method is proposed by Aguirre et al. [78] for VLU versus non-VLU classification from ulcer images. In this study, a pre-trained VGG-19 network was used for classifying the ulcer images in the mentioned two categories. First, a dataset of 300 pictures annotated by a wound specialist was proposed, and data pre-processing and augmentation were conducted before the network training. Then, the VGG-19 network was pre-trained using another

dataset of dermoscopic images. Rostami et al. [79] proposed an end-to-end ensemble DCNN-based classifier to classify the entire wound images into multiple classes, including surgical, diabetic, and venous ulcers. The output classification scores of two classifiers based on patch-wise and image-wise strategies are fed into a Multi-Layer Perceptron to provide a superior classifier. A new dataset of authentic wound images containing 538 images from four different types of wounds was introduced in this research. Sarp et al. [80] classified chronic wounds into four classes (diabetic, lymphovascular, pressure injury, and surgical) by using an explainable artificial intelligence (XIA) approach to provide transparency on the neural network. The dataset contains 8690 wound images collected from the data repository of eKare, Inc. Transfer learning on the VGG16 network was used as the classifier model. The XIA technique can provide explanation and transparency for the wound image classifier and why the model would think a particular class may be present.

Machine and deep learning-based methods are applied for wound tissue classifications. Wannous et al. [81] developed a multi-view strategy for tissue classification (granulation, slough, and necrosis) based on an SVM classifier, relying on a 3-D model where tissue labels were mapped, and classification results are merged. The single view classification results are incorporated into multi-view 3D models in this work. A multi-class classification has been done by Veredas et al. [82] for wound-bed tissue recognition by using three machine learning approaches. They used the k-means clustering method for the segmentation part and then utilized three different classifiers, including neural networks, SVM, and random forest (RF) decision trees for the classification part. They used a dataset of 113 PU wound images. They reported high accuracy rates for the classifiers using SVM and RF trees but the lowest accuracy with neural networks. Veredas et al. [83] classified five types of tissue (necrotic, slough, granulation, healing, and skin) in wound regions using a hybrid approach consisting of neural networks and Bayesian

classifiers. A neural network performed a four-stage-cascaded binary classification followed by a Bayesian classifier. Hazem et al. [84] classified three types of wound tissue (granulation, slough, and necrosis) using an SVM region classifier, with color and texture as input. Using the sequential selection forward (SFS) method, they found 20 descriptors more relevant for the classifier. Mukherjee et al. [85] classified three types of wound tissue (granulation, necrotic, and slough) using Bayesian and SVM classifiers. They considered six types of wounds (burn, DFU, MU, PG, VLU, and PU) from their database. Their experiment showed that SVM provided the best result. Wannous et al. [86] classified four tissue types (granulation, slough, necrosis, and healthy skin) using color and texture features as input to the SVM region classifier. Their experiment showed that the segmentation-driven classification approach was more suitable than pixel-based approach. In their following work [87] they continued segmentation-driven classification of four tissue types by using color and texture features as input to the C-SVM classifier. Their developed tool ensures stability under various lighting conditions, viewpoints, and camera settings.

Zahia et al. [88] proposed a method for tissue analysis of pressure wound images using deep CNNs, trained and tested on a dataset of 22 PU images. For each image, the ground truth labels for pixels were specified by specialists. As a pre-processing task, the authors extracted the ROI from the original wound image, removed the flashlight from the images, and extracted patches of 5×5 pixels from each ROI. These patches and their ground truth labels were fed into the CNN model for training the network to predict three types of wound tissue (granular, slough, and necrotic) from an input image. Based on the results reported, the slough tissue is the most challenging tissue to classify. Rajathi et al. [89] proposed a CNN-based method for tissue classification of varicose ulcer wound images. The dataset used in this study included around 1250 varicose ulcer images collected at a medical college in India and labeled by wound specialists.

Their method is similar to the approach used by Zahia et al. [88], which includes three phases: data pre-processing, active contour segmentation for selecting the wound area from the skin, and CNN-based classification of the input image. The pre-processing step removed the flashlight from the images using thresholding and extracted the ROI. Next, they used a 4-layer CNN to classify wound tissue into four different types (granulation, slough, epithelial, and necrotic) based on the patches of 5×5 pixels generated from the ground truth and the segmented wound image. Nejati et al. [90] proposed a deep learning-based method to analyze chronic wound tissue and classify it into one of the seven classes: necrotic, healthy granulation, slough, infected, unhealthy granulation, hyper granulation, epithelialization. They used a pre-trained AlexNet architecture for feature extraction from the wound tissue and then fed them into an SVM patch-level classifier for classification. The used dataset includes only 350 images. Pasero et al. [91] expanded their segmentation work into tissue classification using a self-organizing map (SOM) and a more extensive dataset.

Blanco et al. [92] proposed a method for dermatological wound image analysis, QTDU, using deep learning models and super-pixel-driven segmentation. This study used a dataset with 217 arterial and venous wound images from lower limbs. The method includes three main steps. The first step comprises labeling the images and constructing and augmenting the super-pixels. They provided 44,893 super-pixels, each tagged with one of four classes: fibrin, granulation, necrosis, and not wound. The second step is data processing and training, in which two deep CNNs (ResNet and InceptionV3) were trained, and six additional layers were added to the end of these deep architectures. The output of the final layer gives the tissue labels. They concluded that the networks pre-trained on the ImageNet dataset were trained faster than the architectures with random weights. In the last step, the pixel-wise wound quantification masks were generated. In addition, the authors said that their proposed method generated better results by using the ResNet

architecture compared to the Inceptionv3 model. Rania et al. [93] performed wound tissue classification into three classes: necrosis, granulation, and slough, by using a superpixel-wise FCN approach. The dataset contains 219 wound images of different types (DFU, VLU, bed sores, etc.). First, they segmented the wound using the U-Net model; then, they generated superpixels using the SLIC method. Each superpixel was given to the network (FCN-32) for prediction, then assigned a class label to each superpixel depending on the dominant color (red, yellow, or black). The authors reported sensitivity, specificity, precision, and DICE score for test evaluation, and the DICE scores for necrosis, slough, and granulation are 59.71%, 77.50%, and 72.71%, respectively.

Though there are several works on wound type and tissue classifications using machine learning and deep learning models, this research is the first study that measures and classifies wound severity from wound images to the best of our knowledge. The dataset collection, dataset preparation, model training process, and experiments done to perform this classification among red, yellow, and green images are discussed in the following sections.

5.2.3 Methodology

5.2.3.1 Dataset Collection

The dataset is collected from the AZH Wound and Vascular Center, Milwaukee, WI, USA. This dataset contains a total of 420 wound images. Among these images, 100 are in the green class, 175 are in the yellow class, and 145 belong to the red class. An expert wound specialist from the AZH center classified the images into these classes. All the images are captured with iPad and DSLR cameras. No specific environmental or illumination condition has been applied during image capturing. These images are further processed and used as training, validation, and test data. Some sample images of each class are shown in Figure 5.9.

(a) Green

(b) Yellow

(c) Red

Figure 5.9: Wound severity database sample images. (a), (b) and (c) rows represent the examples of green, yellow, and red classes, respectively.

5.2.3.2 Dataset Preparation

First, all the images are passed through our developed wound localizer (chapter 5) to select the region of interest (ROIs). We have 723 ROIs from the original images as a single image contains multiple wounds. 80% of the ROIs are used for training and validation, and 20% for testing. The splitting is done carefully so that there should be no overlap between them.

From Figure 5.8, we can see that the peri-wound area has a significant role in the characteristics of the three classes (green, red, and yellow). The peri-wound region is defined as the area of skin that extends beyond the wound edge to a certain amount. To learn the features from the peri-wound area, we applied three zoom-outs on the original ROIs (training and validation set only) produced by our developed YOLOv3 model. The used terms for this experiment are Z0: zoom out zero or original ROIs, Z1: zoom out one with 50 pixels padding on the original ROIs, Z2: zoom out two with 100 pixels padding on the original ROIs, and Z3: zoom out three with 150 pixels padding on the original ROIs. Then, the training set of each Z0 to Z3 is augmented by using horizontal flip, vertical flip, and three rotations (25, 45, and 90 degrees). The dataset preparation process for wound severity classification is shown in Figure 5.10. Table 5.4 shows the dataset splitting and augmentation statistics for wound severity classification.

Table 5.4: Database summary for wound severity classification.

	ROIs			**Augmented ROIs**		
Class	**Train + Validation**	**Test**	**Total**	**Train + Validation**	**Test**	**Total**
Green	154	39	**193**	924	39	**963**
Red	237	60	**297**	1116	60	**1176**
Yellow	186	47	**233**	1422	47	**1469**
Total	**577**	**146**	**723**	**3462**	**146**	**3608**

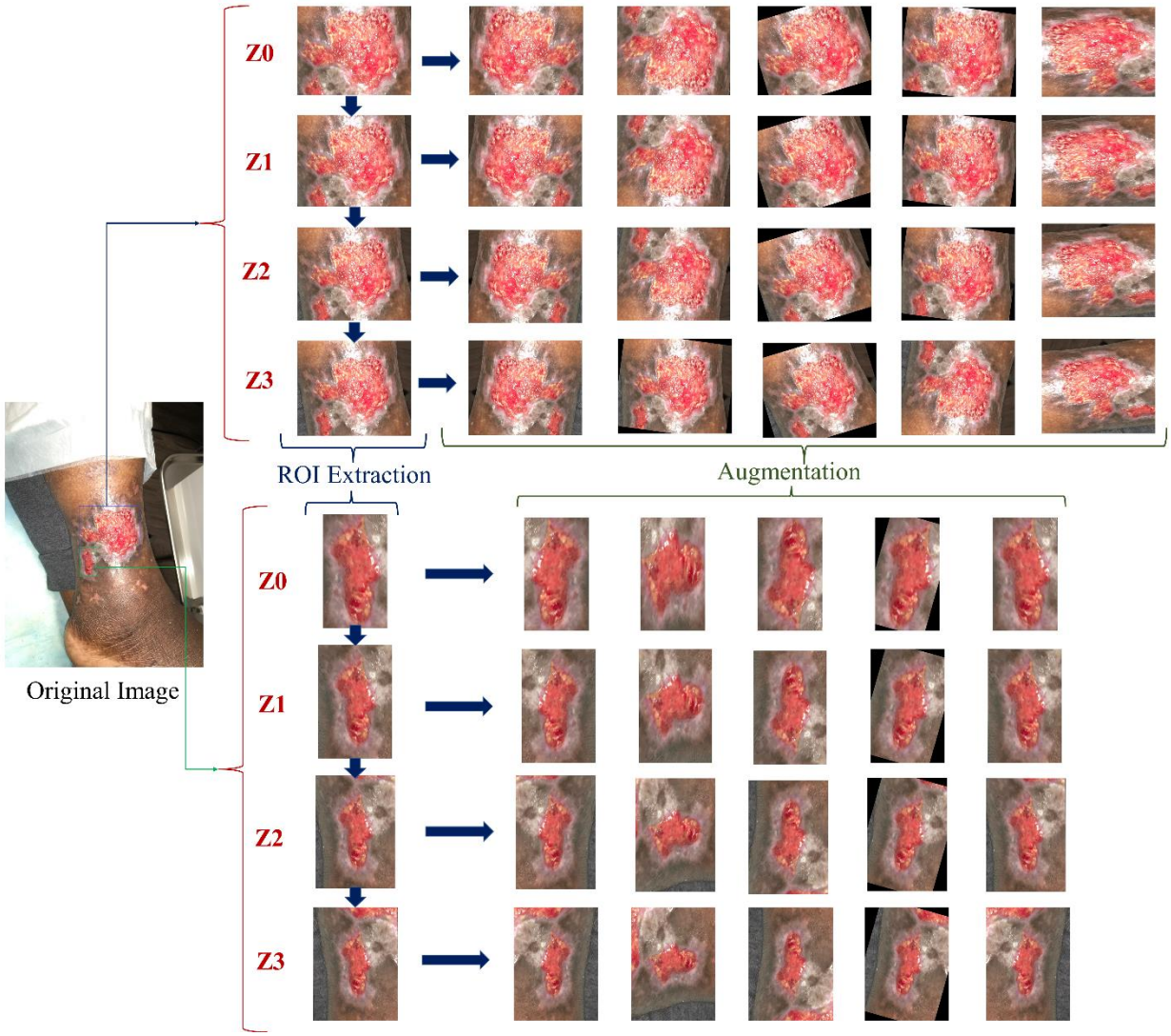


Figure 5.10: Dataset processing steps for wound severity classification.

5.2.3.3 Models

Transfer learning and stacked deep learning models are used to evaluate this dataset. Transfer learning means taking advantage of features learned on one problem and using them in another similar situation. This method is proper when the dataset in hand is small in number to train a full-scale model from scratch, and the memory power is limited to train a vast deep learning model. The most used workflow of transfer learning is: 1) take a previously trained model's layers, 2) freeze the layers, 3) add some new, trainable layers on top of the frozen layers, which will learn

to turn the old features into predictions on a new dataset, and 4) train the new layers on the new dataset [94]. There are 26 deep learning models in Keras Applications [53], among which we choose nine classification models: VGG16 [29], VGG19 [29], InceptionV3 [30], NasNetLarge [35], ResNet50 [31], DenseNet201 [32], Xception [34], MobileNetV2 [36], and InceptionResNetV2 [33]. A brief description of these models is provided in chapter 2. All nine models are pre-trained on ImageNet [95], a huge dataset including more than 14 million general images of 1000 classes.

Stacked models are a mixture of these nine individual models. These stacked models include two different models trained on the same dataset. The stacked models are used in the hope that different models may learn some new features that the other model missed and vice versa. Six stacked models are used, where the individual models are picked based on their standalone results. After images go through each particular network, their outputs are concatenated, and four dense layers are added on top of it to learn about these mixed features, followed by a SoftMax (output) layer.

Finally, nine stacked models are trained, with each having four individual models. These models are different from the previous stacked models because they do not pass the same image through each individual network. Here, four separate networks are trained on the same image's zoom-out images (Z0 to Z3). These nine stacked models are called multi-zoom learning networks. After images go through every four individual networks, their outputs are concatenated, and five dense layers are added on top of it to learn about these mixed features, followed by a SoftMax (output) layer.

All models are written in Python programming language by using Tensorflow.Keras deep learning framework and trained on an Nvidia GeForce RTX 2080Ti GPU platform.

5.2.4 Results and Discussion

5.2.4.1 Results

We use accuracy (equation 2.7) as the performance metric to investigate the classification performance. The confusion matrix of the best model for classifying Z0 images is also provided. Nine transfer learning models and six stacked models are used to classify the original ROIs (Z0). The result of this experiment is shown in Table 5.5.

Table 5.5: Wound severity classification performance on original ROIs (Z0 images).

Model Type	Model	Accuracy
Transfer Learning	VGG16	60.27%
	VGG19	**68.49%**
	InceptionV3	52.05%
	NasNetLarge	56.85%
	ResNet50	41.10%
	DenseNet201	41.10%
	Xception	56.16%
	MobileNetV2	49.32%
	InceptionResNetV2	47.26%
Stacked Models	**M1:** VGG19+NasNetLarge	67.81%
	M2: NasNetLarge + Xception	60.96%
	M3: VGG19+InceptionV3	64.38 %
	M4: VGG16+NasNetLarge	65.75%
	M5: InceptionV3 + Xception	56.16%
	M6: VGG16+InceptionV3	63.70%

Training all these nine individual and six stacked models is expensive in terms of both time and cost. So, we selected five individual models and three stacked models based on their results achieved on the original ROI (Z0) classification to perform classification on zoom-out images (Z1 to Z3). The accuracy of classifying these Z1, Z2, and Z3 images are shown in Table 5.6.

Table 5.6: Wound severity classification performance on zoom-out (Z1, Z2, and Z3) images.

Model Type	**Model**	**Accuracy**		
		Z1	**Z2**	**Z3**
Transfer Learning	VGG16	62.33%	60.96%	55.48%
	VGG19	66.44%	58.22%	61.64%
	InceptionV3	56.16%	56.85%	53.42%
	NasNetLarge	56.85%	61.64%	57.53%
	Xception	56.85%	51.37%	**63.01%**
Stacked Models	**M1:** VGG19+NasNetLarge	67.12%	58.90%	58.22%
	M3: VGG19+InceptionV3	63.70%	56.85%	58.90%
	M4: VGG16+NasNetLarge	**68.49%**	**63.01%**	55.48%

Nine stacked models, each having four individual models, are trained for multi-zoom learning; four individual networks are trained on four zoom-out versions (Z0 to Z3) of the same image. The individual models include all nine transfer learning models. We can not apply the same individual models twice or more because, during concatenation, each layer's name must be unique. Table 5.7 shows the result of this multi-zoom learning. Here, the model's name followed by the zoom-out channel (Z0 to Z3) is trained on the images of that specific zoom-out channel.

Table 5.7: Multi-zoom learning results for wound severity classification.

Stacked Model		Accuracy
M1	**Z0:** VGG19; **Z1:** InceptionV3; **Z2:** NasNet large; **Z3:** Res Net 50	58.90%
M2	**Z0:** Res Net 50; **Z1:** InceptionV3; **Z2:** NasNet large; **Z3:** VGG19	58.22%
M3	**Z0:** Res Net 50; **Z1:** VGG16; **Z2:** NasNet large; **Z3:** InceptionV3	56.85%
M4	**Z0:** Res Net 50; **Z1:** InceptionV3; **Z2:** NasNet large; **Z3:** Xception	52.74%
M5	**Z0:** VGG19; **Z1:** InceptionV3; **Z2:** NasNet large; **Z3:** MobileNetV2	**63.70%**
M6	**Z0:** VGG19; **Z1:** InceptionResNetV2; **Z2:** NasNet large; **Z3:** MobileNetV2	62.33%
M7	**Z0:** DenseNet 201; **Z1:** InceptionResNetV2; **Z2:** NasNet large; **Z3:** MobileNetV2	54.11%
M8	**Z0:** Xception; **Z1:** InceptionResNetV2; **Z2:** DenseNet 201; **Z3:** MobileNetV2	50.68%
M9	**Z0:** VGG19; **Z1:** InceptionResNetV2; **Z2:** Res Net 50; **Z3:** MobileNetV2	60.27%

Finally, we performed three binary classifications on this dataset. The classification performed are green versus yellow, green versus red, and yellow versus red. We used five individual and three stacked models based on their original ROI (Z0) multiclass classification results to perform these three binary classifications. Table 5.8 shows the results of binary classifications on the wound severity database. The binary classifications are performed on the original ROIs (Z0 images).

Table 5.8: Binary classification results on wound severity dataset (Original ROIs).

Model Type	Model	Accuracy		
		Green Vs. Yellow	**Green Vs. Red**	**Yellow Vs. Red**
Transfer Learning	VGG16	71.72%	77.91%	73.83%
	VGG19	76.77%	76.74%	75.70%
	InceptionV3	63.64%	80.23%	64.49%
	NasNetLarge	64.65%	**81.40%**	72.90%
	Xception	70.71%	65.12%	68.22%
Stacked Models	**M1:** VGG19+NasNetLarge	65.66%	74.42%	72.90%
	M3: VGG19+InceptionV3	**78.79%**	77.91%	68.22%
	M4: VGG16+NasNetLarge	63.64%	77.91%	**77.57%**

5.2.4.2 Discussion

From Table 5.5, we can see that the VGG19 model achieves the highest accuracy (68.49%). The second highest accuracy (60.27%) achieved by any individual model is VGG16, the previous version of the VGG19 model. From the stacked models, the highest accuracy (67.81%) is achieved by VGG19 and NasNetLarge combination. Though we expect more accuracy from these stacked models, they can not outperform the VGG19 model. But in general, stacked models did well than individual models. For example, VGG16 and NasNetlarge have individual accuracy of 60.27% and 56.85%; but their stacked model achieved an accuracy of 65.75%.

From Table 5.6, we can see that using zoom-out does not improve the classification performance. Zoom-out one (Z1) with 50 pixels padding images produced the same accuracy as Z0. In the case of Z2 and Z3, we can see a reduction in accuracy value. Though we are using zoom-out images to capture features from the peri-wound regions, increasing the padding size (100 pixels for Z2 and 150 pixels for Z3) of ROIs may capture some unnecessary information like non-skin

regions (including bed sheets, the hand of physicians, wall, floor, etc.), which in turns decreases the accuracy. For Z1 and Z2, the highest accuracy is obtained by the M4 stacked network, which is a combination of VGG16 and NasNetLarge networks. The Xception network achieved the highest accuracy for Z3 images.

From Table 5.7, multi-zoom learning is not helping in the wound severity classification. The highest accuracy was achieved by the M5 model, where Z0 images are passed through the VGG19 network, Z1 images are passed through the InceptionV3 network, Z2 images are passed through the NasNetLarge network, and Z3 images are passed through the MobileNetV2 network. The highest accuracy of multi-zoom learning is 63.70%, almost 5% lower than the accuracy achieved with the original ROIs. As discussed in the above paragraph, the unnecessary information from Z2 and Z3 images may have reduced the classification performance.

From Table 5.8, the highest accuracy achieved for three binary classifications are 78.79%, 81.40%, and 77.57% for green vs. yellow, green vs. red, and yellow vs. red, respectively. The M3 and M4 stacked models achieve the highest accuracies for green vs. yellow and yellow vs. red. For green vs. red, the highest accuracy was acquired by the NasNetlarge network. This result shows that without the middle class (yellow), the green and red images are easier to distinguish.

The confusion matrix for the best model (VGG19) of multiclass wound severity image classification on original ROIs (Z0) is shown in Figure 5.11. The most challenging task was distinguishing between green vs. yellow images and yellow vs. red images from this figure. For this reason, as the middle class, yellow has the highest class-wise recall and lowest class-wise precision. From Figure 5.11, we can see that only 2 green images were misclassified as red, and only one red image was misclassified as green. Some examples of misclassification by the best

model (VGG19) of multiclass wound severity image classification on original ROIs are shown in Figure 5.12.

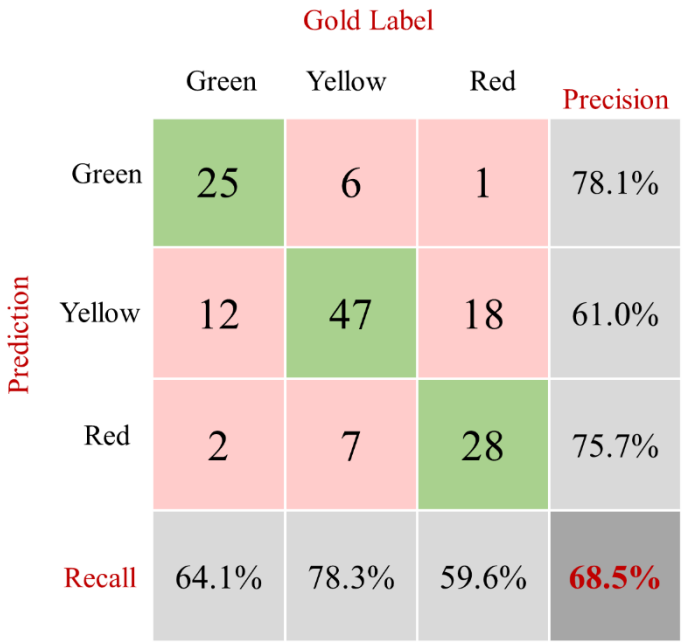


Figure 5.11: Confusion matrix for the best model (VGG19) of multiclass wound severity image classification on the original ROIs.

From Figure 5.12, we can see that it is hard to distinguish between the three classes (green, yellow, and red) from a general human perspective. For example, the second ROI in the top row of Figure 5.12 was annotated as green, but it generally looks like red. Also, the third ROI in the top row of Figure 5.12 was annotated as yellow, but it looks like green. So, from the physician's point of view (who annotated the wound severity dataset), there may be some other features than the standard features (i.e., color, texture, edge, etc.) that are not captured by the convolution and deep neural networks. This statement can also be supported by Figure 5.8. We are now considering only the 2D image features learned through convolutions, but from Figure 5.8, we can see that there are other contributing factors in this classification task (e.g., size, depth, maceration, breakdown, callus tissue, etc.) that must be incorporated to improve the classification performance.

To incorporate these contributing factors (e.g., size, depth, callus tissue, etc.), we may need the output from other wound automation modalities like wound segmentation, wound tissue classification, etc. Finally, we have only 420 wound images, insufficient to develop a sophisticated deep learning model like a wound severity classifier. Also, the collected images on the dataset have lighting issues that hamper the classification performance. The images on the dataset were also not taken from a fixed distance, making the work of wound size calculation much harder. Addressing all these issues and developing an improved and rich dataset will improve the wound severity classification performance by a good margin.

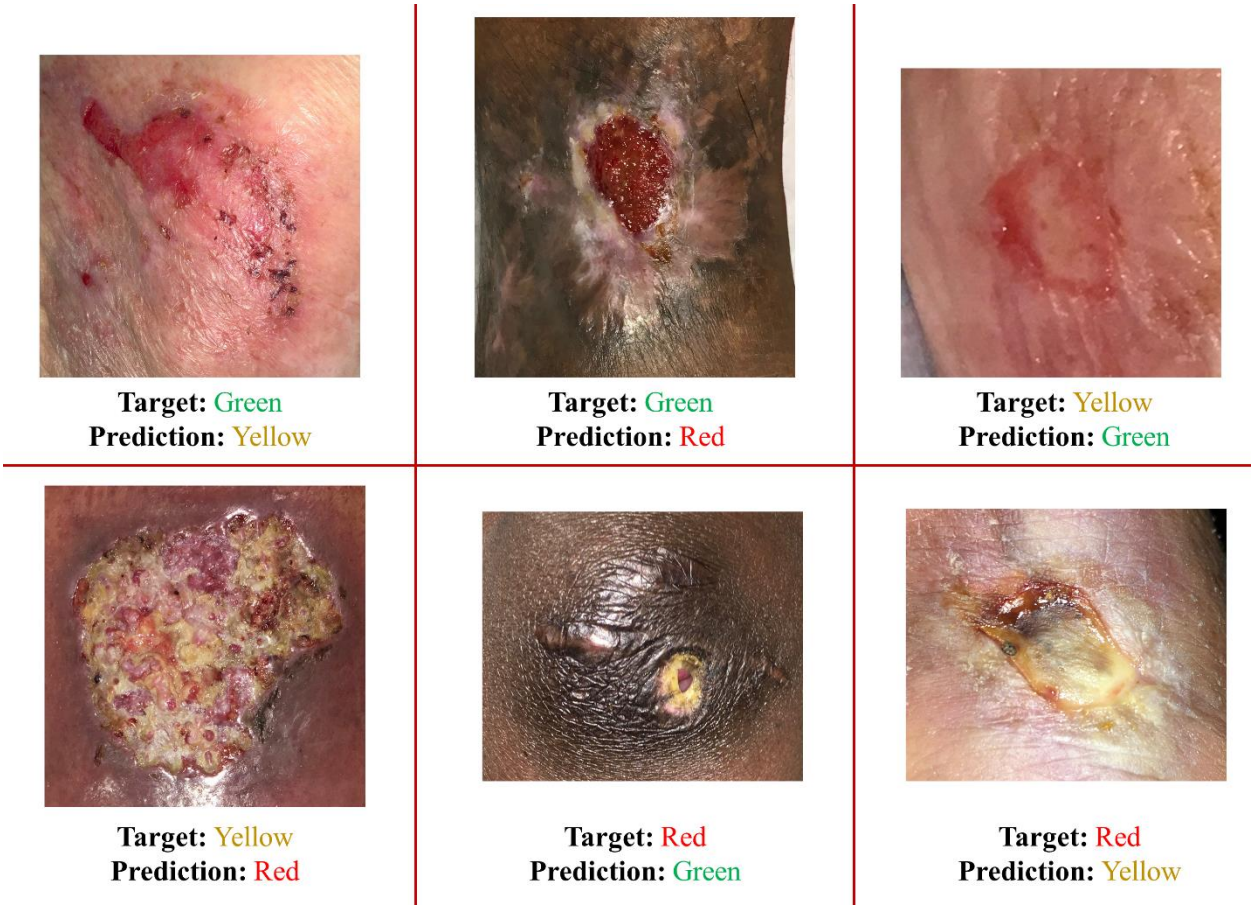


Figure 5.12: Examples of wound severity image misclassifications by the best model (VGG19) for multiclass classification on the original ROIs (Z0).

5.3 Conclusion

In this chapter, two wound image classification problem is discussed. The first problem is burn wound classification. For this task, some deep learning networks are applied to a public dataset, and the evaluation results are promising. However, as the BIP_US dataset has a very small number of images (94 burn images), it is tough to fit a deep neural network because of the hunger for a large amount of data by deep learning networks. This leads to an evaluation result where simpler networks performed best, and deep networks performed worst. Nevertheless, we achieved better performance than the existing state-of-the-art works in burn wound classification.

The second problem is wound severity classification, where several deep learning networks are applied to a novel dataset collected from the AZH wound center. We achieved an accuracy of 68.49% for multiclass classification and 77.57% to 81.40% accuracy for binary classification. The reason behind this poor performance is the relevant factors (size, depth, etc.) which cannot be measured from a 2D image through just convolution. Also, the scarcity of data plays an essential role in model performance. In the future, these relevant factors should be integrated into the model, and the dataset should be polished and increased to improve wound severity classification performance.

Chapter 6

Multi-Modality Wound Classification

6.1 Problem Statement

Wound classification is an essential step of wound diagnosis. An efficient classifier can assist wound specialists in classifying wound types with less financial and time costs and help them decide on an optimal treatment procedure. This chapter discussed a novel deep neural network-based multi-modal classifier, which uses wound images and their corresponding locations to categorize them into multiple classes, including diabetic, pressure, surgical, and venous ulcers. A body map is developed to prepare the location data, which can help wound specialists tag wound locations more efficiently. Three datasets containing images and their corresponding location information are created with the help of wound specialists. The multi-modal network is developed by concatenating the image-based and location-based classifier's outputs with some other modifications.

Though feature-based machine learning and end-to-end deep learning models exist for image-based wound classification, the classification accuracy is limited due to incomplete information considered in the classifiers. The novelty of the present research is to add wound location as a vital feature to obtain a more accurate classification result. Wound location is a standard entry for electronic health record (EHR) document, which many wound physicians utilize for wound diagnosis and prognosis. Unfortunately, these locations are documented manually without specific guidelines, which leads to some inconsistency. In the current work, we developed a body map from which one can visually and accurately select the wound's location. Then, the

wound location is set through the body map for each wound image, and the location is indexed according to the image file name. Finally, the developed classifier is trained with both image (gained through convolution) and location features and produces superior classification performance compared to image-based wound classifiers. A basic workflow of this research is shown in Figure 6.1. The developed wound classifier takes both wound image and location as inputs and outputs the corresponding wound class.

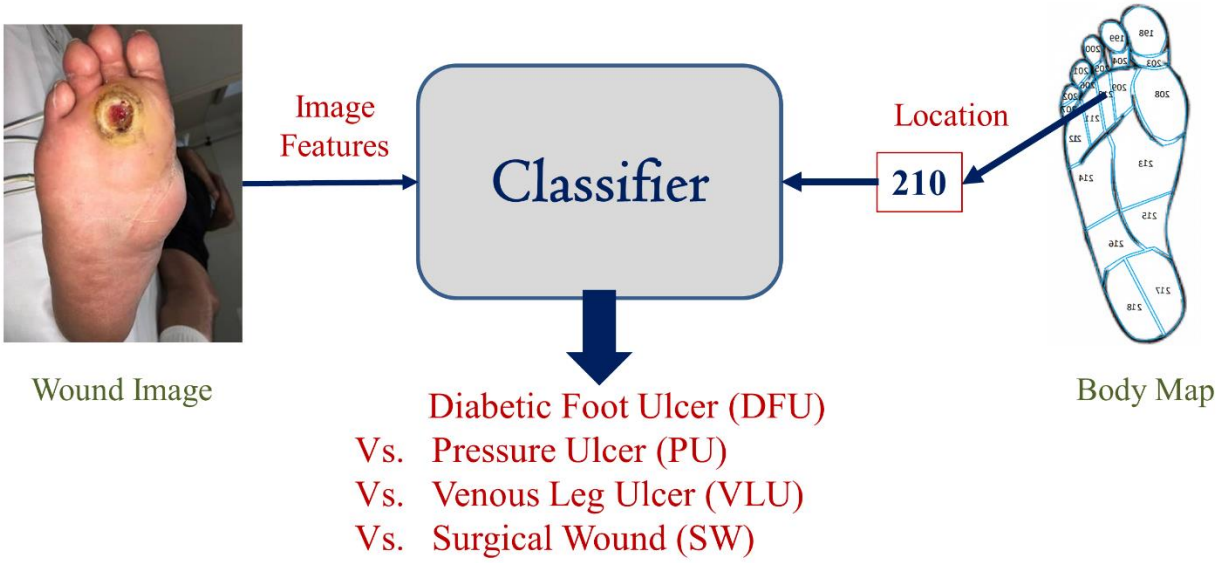


Figure 6.1: Workflow of wound multi-modality classification.

6.2 Related Works

Wound classification includes wound type classification, wound tissue classification, burn depth classification, etc. As we are performing wound type classification, here we discuss existing data-driven wound type classification works. Wound type classification considers different types of wounds and non-wounds (normal skin, background, etc.). Background versus DFU, normal skin

versus PU, and DFU versus PU are binary wound type classification examples. In contrast, DFU versus PU versus VLU is an example of multi-class wound type classification.

Some machine learning-based methods are applied for wound type classification. Goyal et al. [63] used traditional machine learning, deep learning, and ensemble CNN models for binary classification of ischemia versus non-ischemia and infection versus non-infection on DFU images. The authors developed a dataset containing 1459 DFU images that two healthcare professionals labeled. The authors used BayesNet, Random Forest, and Multilayer perceptron for traditional machine learning. Three CNN networks were used as deep-learning approaches. The ensemble CNN contains an SVM classifier that takes three CNN networks' bottleneck features as input. The test evaluation shows that traditional machine learning methods perform the worst, followed by deep-learning networks, while the ensemble CNN performs the best in both binary classifications. Abubakar et al. [70] proposed a machine learning approach to differentiating burn wounds and pressure ulcers. The dataset used in this study includes 29 pressure and 31 burn wound images obtained from the internet and a hospital, respectively. Features were extracted using pre-trained deep architectures like VGG-face, ResNet101, and ResNet152 from the images and then fed into an SVM classifier to classify the images into burn or pressure wound classes. Several experiments were conducted, including binary classification (burn or pressure) and 3-class classification (burn, pressure, and healthy skin).

Deep learning-based methods are also applied for wound type classifications. A novel CNN architecture named DFUNet was developed by Goyal et al. [58] for binary classification of healthy skin and DFU skin. A dataset of 397 wound images was presented, and data augmentation techniques were applied to increase the number of images. The proposed DFUNet utilized concatenating the outputs of three parallel convolutional layers with different filter sizes. Shenoy

et al. [59] proposed a CNN-based method for binary classification of wound images. This study used a dataset of 1335 wound images collected via smartphones and the internet. The authors considered nine different labels (wound, infection (SSI), granulation tissue, fibrinous exudates, open wound, drainage, steri strips, staples, and sutures) for the dataset, where for each label, two subcategories (positive and negative) were considered. The authors used a modified VGG16 network named WoundNet as the classifier, pre-trained using the ImageNet dataset. In addition, the researchers created another network called Deepwound, an ensemble model that averages the results of three individual models. A binary patch classification of normal skin versus abnormal skin (DFU) is performed by Alzubaidi et al. [60] with a novel deep convolutional neural network named DFU_QUTNet. First, the authors introduced a new dataset consisting of 754-foot images from a diabetic hospital center in Iraq. The proposed network is a deep architecture with 58 layers, including 17 convolutional layers. The performance of their proposed method was compared with those of other deep CNNs like GoogLeNet, VGG16, and AlexNet.

A CNN-based method is proposed by Aguirre et al. [78] for VLU versus non-VLU classification from ulcer images. In this study, a pre-trained VGG-19 network was used for classifying the ulcer images in the mentioned two categories. First, a dataset of 300 pictures annotated by a wound specialist was proposed, and data pre-processing and augmentation were conducted before the network training. Then, the VGG-19 network was pre-trained using another dataset of dermoscopic images. Rostami et al. [79] proposed an end-to-end ensemble DCNN-based classifier to classify the entire wound images into multiple classes, including surgical, diabetic, and venous ulcers. The output classification scores of two classifiers based on patch-wise and image-wise strategies are fed into a Multi-Layer Perceptron to provide a superior classifier. A new dataset of authentic wound images containing 538 images from four different types of wounds was

introduced in this research. Sarp et al. [80] classified chronic wounds into four classes (diabetic, lymphovascular, pressure injury, and surgical) by using an explainable artificial intelligence (XIA) approach to provide transparency on the neural network. The dataset contains 8690 wound images collected from the data repository of eKare, Inc. Transfer learning on the VGG16 network was used as the classifier model. The XIA technique can provide explanation and transparency for the wound image classifier and why the model would think a particular class may be present.

Though some wound type classification works from wound images exist, there is no automated wound classification work based on the wound location feature, to the best of our knowledge. This research is the first work that incorporates wound location for automatic wound type classification and proposes a multi-modal network that uses both wound image features and location features to classify a wound.

6.3 Methodology

6.3.1 Dataset Collection

In this research, three different datasets are used for the robustness and reliability of the model performance. A brief discussion of these datasets is given below:

6.3.1.1 AZH Dataset

AZH dataset is collected over a two-year clinical period at the AZH Wound and Vascular Center in Milwaukee, Wisconsin. The dataset includes 730 wound images in .jpg format. The images are of various sizes, where the width ranging from 320 to 700 pixels and the height ranging from 240 to 525 pixels. These images contain four different wound types: venous, diabetic, pressure, and surgical. iPad Pro (software version 13.4.1) and a Canon SX 620 HS digital camera

are used to capture the images, and labeling is done by a wound specialist from the AZH Wound and Vascular Center. Each image is taken from a separate patient for most images in our dataset. But there are a few cases where multiple photos were taken from the same patient at different body sites or various healing stages. For the latter case, the wound shapes are different, so they are considered separate images.

6.3.1.2 Medetec Dataset

Medetec wound database [64] contains free stock images of all types of open wounds. We randomly collect 358 images from these three categories: diabetic, pressure, and arterial and venous leg ulcers. The arterial and venous leg ulcer images are not separated in the Medetec database, so we consider them in the same category. This dataset does not contain any surgical wound images. All the images are in .jpg format, where the weight varies from 358 to 560 pixels, and the height varies from 371 to 560 pixels.

6.3.1.3 AZHMT Dataset

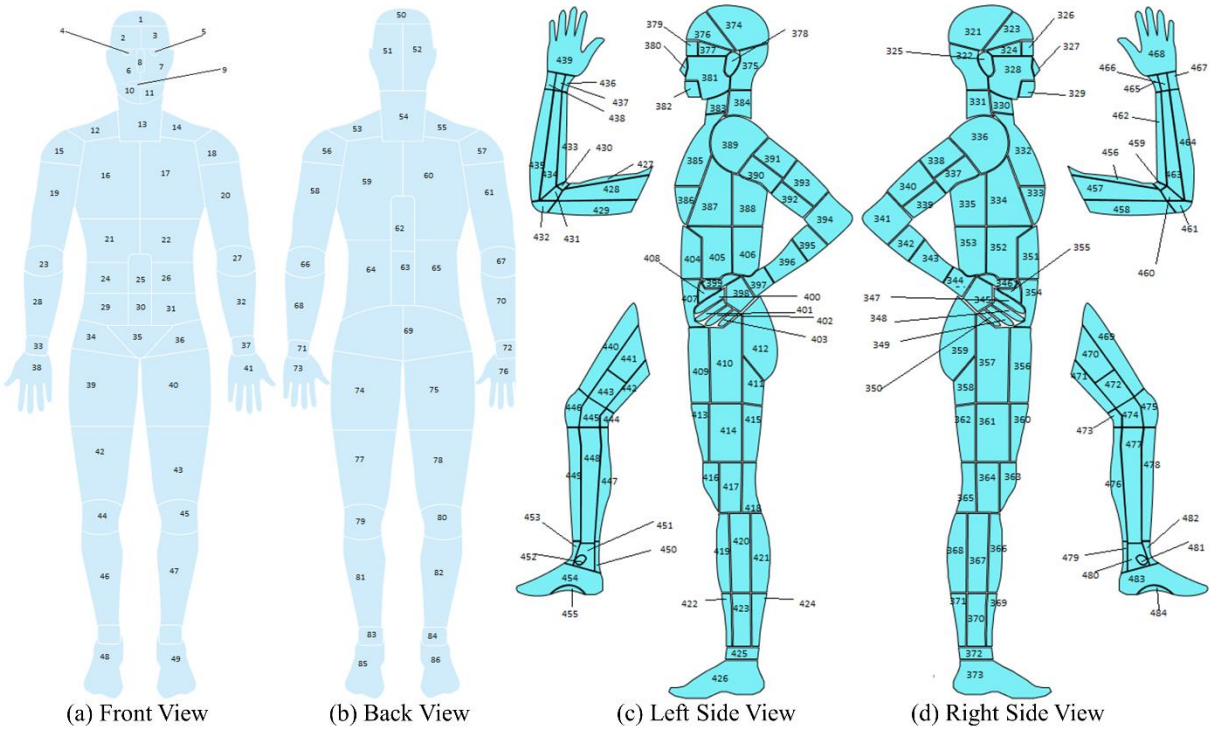
This dataset is the mixer of all the images from the AZH and Medetec datasets. This dataset contains 1088 wound images in .jpg format. AZHMT includes four wound classes: diabetic, pressure, surgical, and arterial + venous leg ulcers. The width of these images varies from 320 to 700 pixels, and the height ranges from 240 to 560 pixels. AZHMT dataset is created for the reliability and robustness testing of the developed model.

6.3.2 Body Map for Location

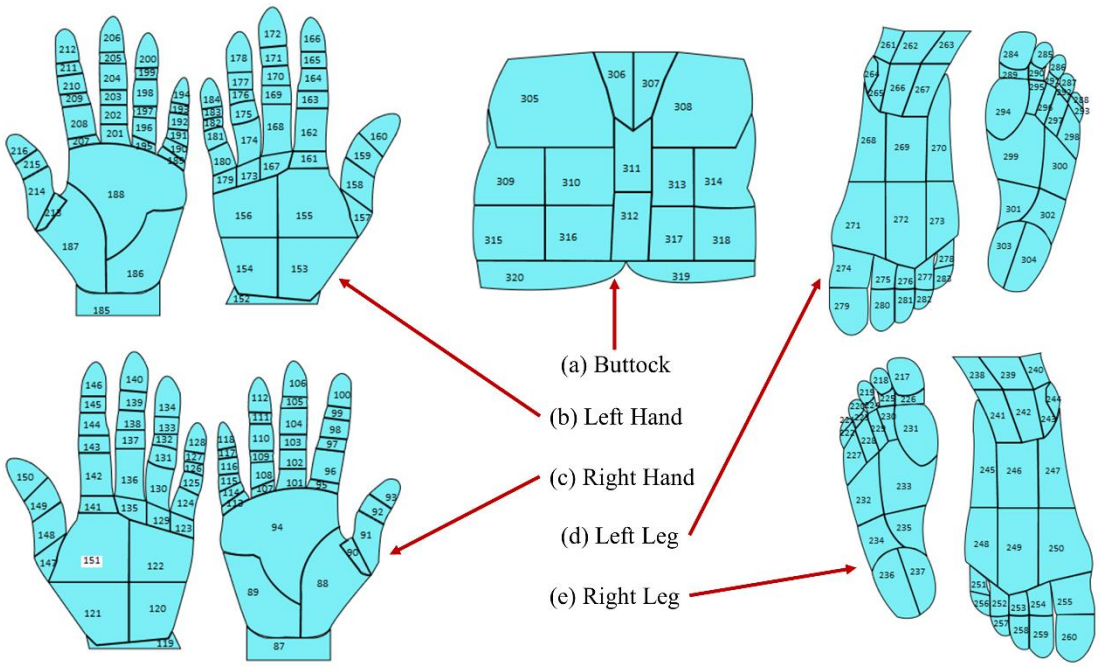
A body map is a labeled, simplified, and symbolic diagram of the entire body of the person, which should be phenotypically right [96]. Medical practitioners use body maps to locate bruises, wounds, or body breakage on a patient's body. Doctors use body maps to analyze the location of a given infection in patients [97]. A detailed body map helps doctors determine which other part of the body to be cautious about during the wound's rehabilitation process.

Table 6.1: Examples of body locations and their corresponding mapping.

Location	Reference Number
Left Hand Front	
Left Dorsal Wrist	152
Left Proximal Lateral Dorsal Hand	153
Right Leg Bottom	
Right Distal Plantar First Toe	217
Right Distal Lateral Mid Plantar Foot	232
Buttock	
Left Posterior Lower Back	305
Superior Gluteal	311



(i) Full Body View

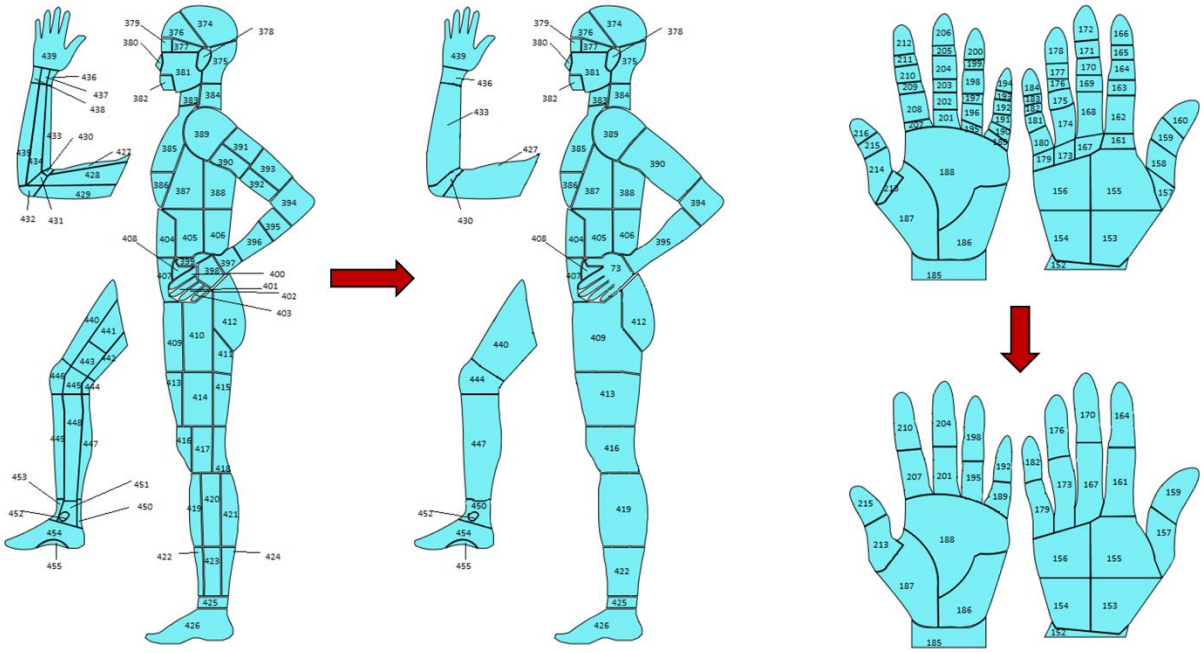


(ii) Detailed Body View

Figure 6.2: Body Map for wound location selection.

A simplified body map with 484 total parts is designed to avoid the body map's complexity. The body map is prepared using PaintCode [98]. The initial reference to the body map is obtained from [99]–[101]. The ground truth diagram for the design is based on the Original Anatomy Mapper [102]. Each label and outline are directly paired with the labeling provided by the anatomy mapper [102]. To avoid the extreme complexity of drawing every detailed feature of the body map, a total of 436 feature or region is pre-selected and approved by wound professionals at the AZH wound and vascular center. The developed body map is shown in Figure 6.2. Here each number represents a location. A few examples of the locations and their corresponding numbers are shown in Table 6.1.

Through experiments, we observe that our number of data (images) is deficient regarding the different wound types and locations, which leads to very few data points per class. To maintain the reliability of the experiment, the body map is further simplified by merging different sections of our developed body map. For example, body locations 436, 437, and 438 are merged and referenced as 436; similarly, body locations 390, 391, 392, and 393 are combined and referenced as 390; and so on. With this simplification, 161 location points are removed from our developed body map, and the total number of locations is decreased from 484 to 323. This makes our location classifier predict more realistic results, making the whole experiment reliable. More details are discussed in the “Selecting Best Experimental Setup” section. Some examples of simplified body maps are shown in Figure 6.3. Our developed original body map is discussed here because, with the increment of the number of images, we will use this body map with 484 body locations in the future. For this research, we used the simplified body map containing 323 locations.



(a) Left Side View Simplification (b) Left Hand Simplification

Figure 6.3: Body Map simplification.

6.3.3 Dataset Processing

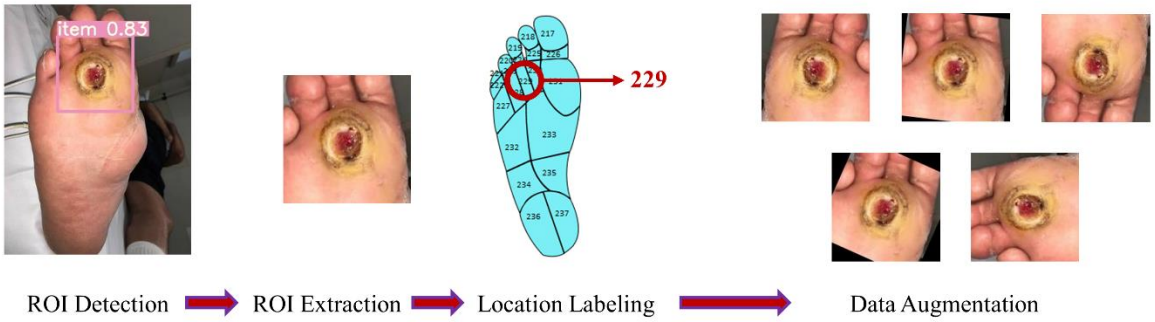
All datasets go through three steps: region of interest (ROI) cropping, location labeling, and data augmentation. First, single or multiple ROIs have cropped automatically from each image using our developed wound localizer (chapter 5). The extracted ROIs are rectangular, but their height and weight are different depending on the wound size. Then all the ROI's locations are labeled by a wound specialist. The location labeling is done by using our developed body map. As our body map represents each location with a unique number, each ROI is tagged with a location number for model training. Finally, rotation and flipping augmentations are used for each ROI to increase the data numbers. A total of five augmentations are applied to each ROI: horizontal and vertical flip, 25, 45, and 90-degree rotations. As wound location does not change with image

augmentation, the location number is repeated for each augmented image. The dataset processing step is illustrated in Figure 6.4.

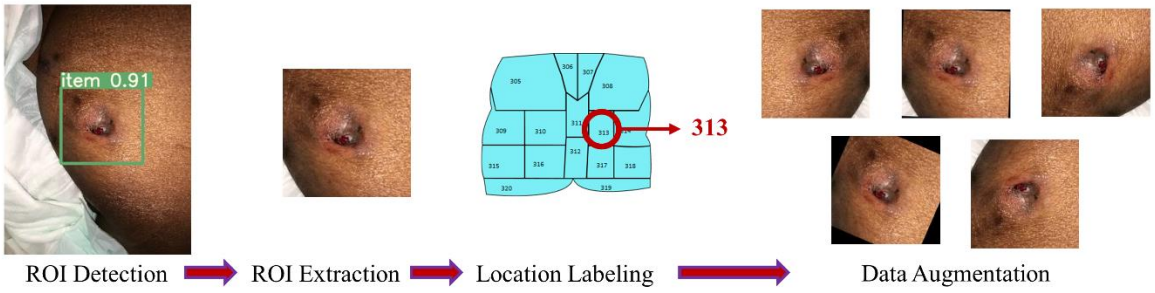
Each dataset (ROI) is divided into 60% training, 15% validation, and 25% test sets. First, the 25% test set is created from a random selection of the wound images to ensure no overlap between training and test sets. The validation set is also created randomly during the time of training. Next, the 75% training and validation datasets are augmented, while test images do not go through data augmentation. Two non-wound classes named normal skin and background are created by manually cropping corresponding ROIs from the original images. A wound specialist does the location tagging for healthy skin. As the background ROIs do not represent any location of our developed body map, each ROI is tagged with a location number '-1'. Table 6.2 shows the number of images of all three datasets. All the six classes, diabetic, venous, arterial + venous, pressure, surgical, background, and normal skin, are represented with D, V, A+V, P, S, BG, and N, respectively.

Table 6.2: Description of all datasets for wound multimodality classification.

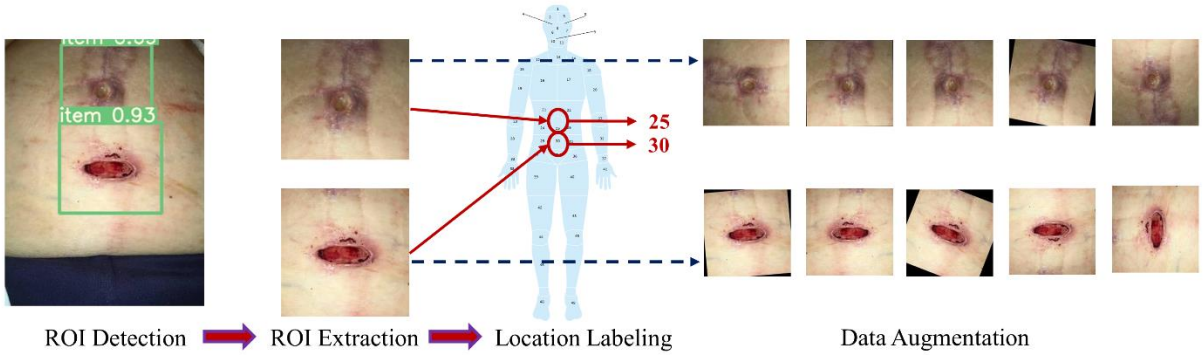
Dataset:	**AZH**			**Medetec**			**AZHMT**		
Class	**Training & Validation**	**Test**	**Total**	**Training & Validation**	**Test**	**Total**	**Training & Validation**	**Test**	**Total**
Background (BG)	450	25	475	0	0	0	450	25	475
Normal Skin (N)	450	25	475	0	0	0	450	25	475
Diabetic (D)	834	46	880	330	19	349	1164	65	1229
Pressure (P)	600	34	634	822	46	868	1422	80	1502
Surgical (S)	732	42	774	0	0	0	732	42	774
Venous (V)	1110	62	1172	0	0	0	0	0	0
Arterial + Venous (A+V)	0	0	0	456	25	481	1566	87	1653
Total	**4176**	**234**	**4410**	**1608**	**90**	**1698**	**5784**	**324**	**6108**



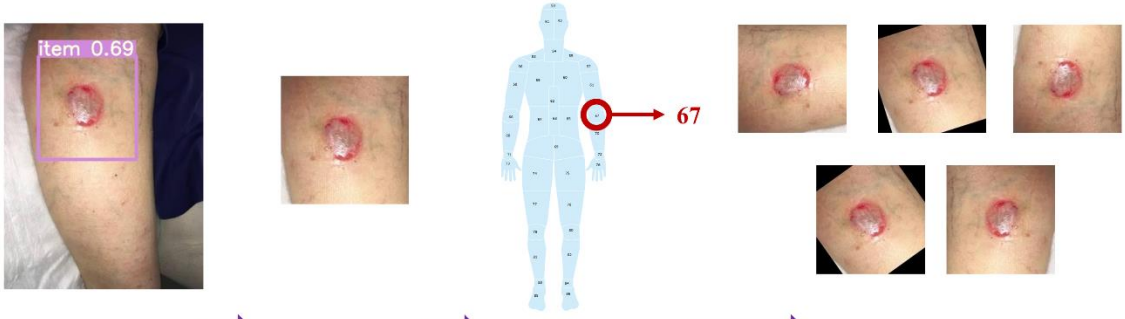
(a) Diabetic Foot Ulcer



(b) Pressure Ulcer



(c) Surgical Wound



(d) Venous Ulcer

Figure 6.4: Dataset processing steps for wound type classification.

6.3.4 Model

As our dataset contains both image and categorical (wound location) data, Keras Functional API [103] is used to develop a network that can handle multiple inputs and mixed data. The Functional API is more flexible than the Sequential API, which can control models with non-linear topology, shared layers, and multiple inputs or outputs. Considering a deep learning model as a directed acyclic graph (DAG) of layers, the functional API is a way to build graphs of layers. Figure 6.5 shows the architecture of our wound-type classification network.

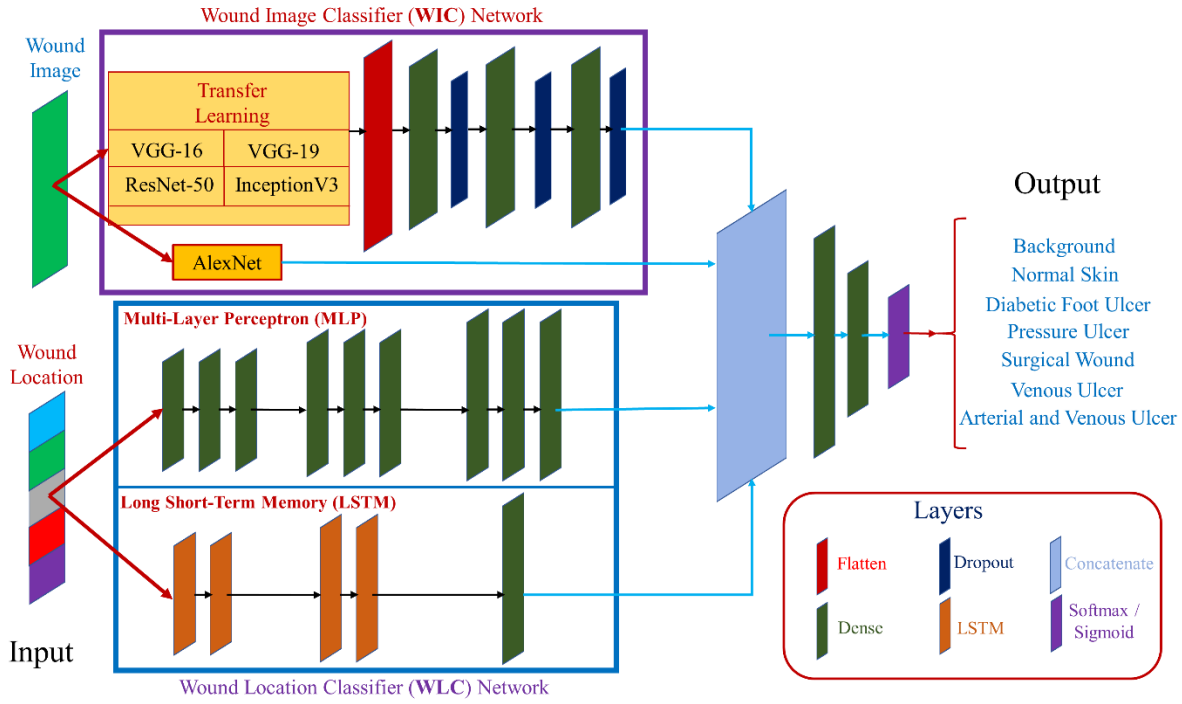


Figure 6.5: Wound Multimodality Classifier (WMC) network architecture.

Two separate neural networks for each data type are used to work with both image and location data. These networks are then considered input branches, and their outputs are combined into a final neural network. We address the image network as Wound Image Classifier (WIC) network, the location network as Wound Location Classifier (WLC) network, and the combined

network as Wound Multimodality Classifier (WMC) network. The output of this WMC network is the probability of the wound class.

The multi-modal network (WMC) must arrange the data in the correct order. The output for the image and location data must be consistent, so the final combined (WMC) neural network must be fed with the right ordered data simultaneously. For example, to train the WMC network properly, we must give the output of the WIC network for the 148th DFU image and the output of the WLC network for the 148th DFU wound's location as the input at the same time to the WMC network. If the data are not ordered correctly, the WMC network may have the WIC network's output for the 148th DFU image and the WLC network's output for the 55th PU wound's location as input at the same time, which will lead to a wrong classification. This arrangement is taken care of by giving each ROI a unique index number and tagging the corresponding location to that index number.

6.3.4.1 Wound Image Classifier (WIC) Network

The wound image classifier (WIC) network is built upon transfer learning, except the AlexNet [28]. Among the 26 deep learning models in Keras Applications [53], we choose four top-rated classification models: VGG16 [29], VGG19 [29], InceptionV3 [30], and ResNet50 [31]; and take their previously trained layers to apply transfer learning. Except for the top layer, all the layers are frozen for all these four models, and three Dense layers with dropout layers are added (Figure 6.5, top WIC box) for training on our wound datasets. All three Dense layers contain 512 trainable neurons, all having the ReLU activation. The AlexNet [28] is implemented following the

original architecture. The output layer is added with either softmax or sigmoid layer for multi-class or binary-class classification for all the models, respectively.

6.3.4.2 Wound Location Classifier (WLC) Network

The Wound Location Classifier (WLC) network can classify wound locations using either a Multi-Layer Perceptron (MLP) or Long Short-Term Memory (LSTM) network. A brief description of MLP and LSTM is provided in chapter 2. As the location data is categorical, we used one-hot encoding to represent the data, representing each input to the WLC network as a one-hot vector. The WLC network handles only one categorical data (location), for which the architecture of the network is kept simple. With a deeper network, the accuracy does not improve (sometimes decreases), and resources (time and memory) become expensive. The MLP network contains nine Dense layers, all having the ReLU activation. The first three layers contain 128 neurons, the following three layers contain 256 neurons, and the last three layers contain 512 neurons (Figure 6.5, middle MLP box). The LSTM contains four LSTM layers, followed by a Dense layer, all having the ReLU activation. The first two layers contain 32 neurons, followed by two LSTM layers having 64 neurons each, and finally, the Dense layer contains 512 neurons (Figure 6.5, bottom LSTM box). The output layer is added with either softmax or sigmoid layer for multi-class or binary-class classification for all the models, respectively.

6.3.4.3 Wound Multimodality Classifier (WMC) Network

As discussed earlier, the Wound Multimodality Classifier (WMC) network is designed using Keras Functional API, which can predict the wound classes based on both wound image and location information. At first, the image data goes through the WIC network, the location data goes through the WLC network, and the networks' outputs are concatenated. Then, two Dense layers are added after concatenation to learn from the merged features. These Dense layers contain 512 and 256 neurons, respectively. Finally, the output layer is added with either a softmax or sigmoid layer for multi-class or binary-class classifications.

6.4 Experimental Setup

Lots of experiments are performed with different setups. Classification between D vs. V, D vs. S, N vs. D, etc. are some examples of binary classification, and D vs. P vs. S, BG vs. N vs. S vs. V, BG vs. N vs. D vs. P vs. S vs. V, etc. are some examples of multi-class classification. In the WMC network, all combinations of the WIC and WLC networks (AlexNet+MLP, AlexNet+LSTM, ResNet50+MLP, VGG16+LSTM, etc.) are applied for the four wound class classification (D vs. P vs. S vs. V) on the AZH dataset. Based on the results (discussed later), the best two combinations are applied for the other multi-modal classifications.

All the models are written in Python programming language using the Keras deep learning framework and trained on an Nvidia GeForce RTX 2080Ti GPU platform. All models are trained for 250 epochs with a batch size of 25, a learning rate of 0.001, and an Adam optimizer. Two callbacks are used with the best validation accuracy and the best combination of validation and

training accuracy saving. For multi-class classification and binary class classification, *sparse_categorical_crossentropy* and *binary_crossentropy* loss functions are used, respectively.

We use accuracy (equation 2.7) as the performance metric to investigate the classification performance. We also use precision, recall, and f1-score (equation 2.2, 2.3, and 2.4) as performance metrics to evaluate binary classifications.

6.5 Results

6.5.1 Selecting Best Experimental Setup

Four wound class classification (D vs. P vs. S vs. V) on the AZH dataset is chosen to select the best combinations for the WMC network. This classification is the most challenging task, as there are no normal skin (N) or background (BG) images in the experiment. This experiment is done with our original developed body map, which contains 484 locations. Table 6.3 shows the results of this experiment. We also present the results on the original dataset (without any augmentation) for this experiment to show the effect (improvement) of data augmentation. The performances of MLP and LSTM are similar on the WLC network, and the VGG16 and VGG19 perform best on the WIC network. Their combinations: VGG16+MLP, VGG19+MLP, VGG16+LSTM, and VGG19+LSTM, also work best for the WMC network. The performance of AlexNet+MLP, AlexNet+LSTM, ResNet50+MLP, and ResNet50+LSTM are very poor. The InceptionV3+MLP and InceptionV3+LSTM performances are not good enough to apply to all the experiments. Running all these combinations for many experiments is also expensive (both with time and memory). So, we apply the best four combinations (VGG16+MLP, VGG19+MLP, VGG16+LSTM, and VGG19+LSTM) for all the remaining experimental setups.

Table 6.3: Results of four wound class classifications (D vs. P vs. S vs. V) on the AZH dataset with original body map.

Input	Model	Original Dataset	Augmented Dataset
		Accuracy	Accuracy
Location	MLP	66.30%	71.74%
	LSTM	66.85%	72.28%
Image	AlexNet	35.33%	37.50%
	VGG16	65.76%	71.73%
	VGG19	56.52%	63.04%
	InceptionV3	51.09%	56.52%
	ResNet50	33.70%	33.70%
Image + Location	AlexNet+MLP	55.43%	61.41%
	VGG16+MLP	77.17%	78.26%
	VGG19+MLP	62.50%	72.28%
	InceptionV3+MLP	61.41%	70.11%
	ResNet50+MLP	63.04%	66.85%
	AlexNet+LSTM	58.15%	66.85%
	VGG16+LSTM	72.83%	**79.35%**
	VGG19+LSTM	71.20%	76.63%
	InceptionV3+LSTM	64.67%	69.02%
	ResNet50+LSTM	33.70%	34.79%

The same four wound class classification (D vs. P vs. S vs. V) on the AZH dataset is done with the simplified body map containing 323 locations. Table 6.4 shows the comparison of this experiment's result with the previous result (shown in Table 6.3). The image classifier (WIC) does not affect the change in the body map, for which it is excluded from Table 6.4. Therefore, we use the simplified body map for all the remaining experiments with improved accuracy in all models.

Table 6.4: Results of four wound class classifications (D vs. P vs. S vs. V) on the AZH dataset with simplified body map.

Input	Model	Accuracy with Original Body Map	Accuracy with Simplified Body Map
Location	MLP	71.74%	74.46 %
	LSTM	72.28%	73.37%
Image + Location	VGG16+MLP	78.26%	**81.52%**
	VGG19+MLP	72.28%	78.80%
	VGG16+LSTM	**79.35%**	80.43%
	VGG19+LSTM	76.63%	79.89%

6.5.2 Experiment on AZH Dataset

A classification between all the classes is performed on the AZH dataset. Table 6.5 shows the results of this six-class classification (BG vs. N vs. D vs. P vs. S vs. V). We achieve the highest accuracy of 82.48% with the multi-modal (WMC) network using the VGG19+MLP combination, where the highest accuracies reach from WLC and WIC networks are 67.52% and 75.64% using LSTM and VGG16 networks, respectively.

Table 6.5: Six-class classification (BG vs. N vs. D vs. P vs. S vs. V) results on the AZH Dataset.

Input	Model	Accuracy
Location	MLP	64.96%
	LSTM	67.52%
Image	VGG16	75.64%
	VGG19	64.96%
Image + Location	VGG16+MLP	79.49%
	VGG19+MLP	**82.48%**
	VGG16+LSTM	79.91%
	VGG19+LSTM	72.22%

Four five-class classifications are performed on the AZH dataset. The classifications are 1) BG vs. N vs. D vs. P vs. V, 2) BG vs. N vs. D vs. S vs. V, 3) BG vs. N vs. D vs. P vs. S, and 4) BG vs. N vs. P vs. S vs. V. We achieve the highest accuracy of 86.46%, 91.00%, 83.14%, and 86.17% for classification number 1), 2), 3), and 4), respectively. The highest accuracy is achieved with the multi-modal (WMC) networks in all four classifications. Table 6.6 shows the detailed results of these classifications.

Table 6.6: Results of four five-class classifications on the AZH Dataset.

Classifications		**BG-N-D-P-V**	**BG-N-D-S-V**	**BG-N-D-P-S**	**BG-N-P-S-V**
Input	**Model**	**Accuracy**			
Location	MLP	67.71%	75.00%	59.30%	69.68%
	LSTM	68.75%	72.00%	59.30%	71.81%
Image	VGG16	69.79%	70.50%	64.53%	75.53%
	VGG19	76.56%	74.50%	67.44%	72.34%
Image + Location	VGG16+MLP	**86.46%**	85.00%	**83.14%**	84.04%
	VGG19+MLP	85.42%	86.50%	77.33%	**86.17%**
	VGG16+LSTM	84.38%	**91.00%**	77.33%	77.13%
	VGG19+LSTM	78.65%	89.50%	73.26%	75.00%

Six four-class classifications are performed on the AZH dataset, along with one wound class classification (shown in Table 6.3 and Table 6.4). The classifications are: 1) BG vs. N vs. D vs. V, 2) BG vs. N vs. P vs. V, 3) BG vs. N vs. S vs. V, 4) BG vs. N vs. D vs. P, 5) BG vs. N vs. D vs. S, and 6) BG vs. N vs. P vs. S. We achieve the highest accuracy of 95.57%, 92.47%, 94.16%, 89.23%, 91.30%, and 85.71% for classification number 1), 2), 3), 4), 5), and 6), respectively. The highest accuracy is achieved with the multi-modal (WMC) networks in all six classifications. Table 6.7 shows the detailed results of these classifications.

Table 6.7: Results of six four-class classifications on the AZH Dataset.

Classifications		BG-N-D-V	BG-N-P-V	BG-N-S-V	BG-N-D-P	BG-N-D-S	BG-N-P-S
Input	**Model**	**Accuracy**					
Location	MLP	76.58%	73.29%	77.27%	65.38%	71.74%	69.04%
	LSTM	78.48%	76.03%	83.12%	64.62%	73.91%	67.46%
Image	VGG16	93.67%	89.73%	87.66%	82.31%	77.54%	83.33%
	VGG19	89.87%	86.99%	88.31%	80.00%	81.88%	83.33%
Image + Location	VGG16+ MLP	94.30%	91.78%	**94.16%**	86.15%	86.96%	**85.71%**
	VGG19+ MLP	**95.57%**	91.78%	92.86%	86.92%	**91.30%**	81.75%
	VGG16+ LSTM	89.87%	**92.47%**	90.91%	86.15%	84.78%	83.33%
	VGG19+ LSTM	94.30%	89.04%	88.89%	**89.23%**	85.51%	83.33%

Four three-wound-class classifications are performed on the AZH dataset. The classifications are 1) D vs. S vs. V, 2) P vs. S vs. V, 3) D vs. P vs. S, and 4) D vs. P vs. V. We achieve the highest accuracy of 92.00%, 85.51%, 72.95%, and 84.51% for classification number 1), 2), 3), and 4), respectively. The highest accuracy is achieved in all four wound-class classifications with the multi-modal (WMC) networks. Table 6.8 shows the detailed results of these classifications.

Table 6.8: Results of four three-wound-class classifications on the AZH Dataset.

Classifications		D-S-V	P-S-V	D-P-S	D-P-V
Input	**Model**	**Accuracy**	**Accuracy**	**Accuracy**	**Accuracy**
Location	MLP	81.33%	82.61%	65.57%	78.87%
	LSTM	82.00%	80.43%	68.85%	78.87%
Image	VGG16	74.67%	68.12%	61.48%	76.06%
	VGG19	76.00%	70.23%	58.20%	68.31%
Image + Location	VGG16+MLP	85.33%	**85.51%**	70.49%	80.28%
	VGG19+MLP	**92.00%**	82.61%	71.31%	**84.51%**
	VGG16+LSTM	80.67%	81.88%	**72.95%**	83.10%
	VGG19+LSTM	87.33%	68.12%	67.21%	**84.51%**

Finally, ten binary classifications are performed on the AZH dataset. The classifications are: 1) N vs. D, 2) N vs. P, 3) N vs. S, 4) N vs. V, 5) D vs. P, 6) D vs. S, 7) D vs. V, 8) P vs. S, 9) P vs. V, and 10) S vs. V. We achieve highest accuracy of 100%, 98.31%, 98.51%, 100%, 85.00%, 89.77%, 94.44%, 89.47%, 90.63%, and 97.12% for classification number 1), 2), 3), 4), 5), 6), 7), 8), 9), and 10), respectively. The highest accuracy is achieved with the multi-modal (WMC) networks in all binary classifications. Table 6.9 shows the detailed results of these binary classifications. The precision, recall, and f1-score for all the best models (according to accuracy) are also calculated and shown in Table 6.10.

Table 6.9: Accuracy of ten binary classifications on the AZH Dataset.

Classifications		**N-D**	**N-P**	**N-S**	**N-V**	**D-P**	**D-S**	**D-V**	**P-S**	**P-V**	**S-V**
Input	**Model**	**Accuracy (in %)**									
Location	MLP	78.9	64.4	74.6	78.2	78.8	87.5	89.8	73.7	87.5	93.3
	LSTM	77.5	43.4	76.1	78.2	78.8	81.8	57.4	73.7	85.4	93.3
Image	VGG16	98.6	96.6	97.0	98.9	81.3	79.6	88.0	77.6	84.4	84.6
	VGG19	98.6	**98.3**	97.0	98.9	71.3	80.7	88.0	73.7	86.5	86.5
Image + Location	VGG16 +MLP	97.2	96.6	**98.5**	98.9	80.0	**89.8**	**94.4**	**89.5**	88.5	94.2
	VGG19 +MLP	95.8	94.9	97.0	98.9	80.0	84.1	92.6	80.3	**90.6**	**97.1**
	VGG16 +LSTM	97.2	96.6	95.5	98.9	83.8	80.7	**94.4**	76.3	83.3	84.6
	VGG19 +LSTM	**100**	**98.3**	97.0	**100**	**85.0**	77.3	88.9	71.1	82.3	79.8

Table 6.10: Precision, Recall, and F1-Scores of the best models of ten binary classifications on the AZH Dataset.

Classifications	Best Model(s)	Precision	Recall	F1-Score
N-D	VGG19+LSTM	100%	100%	100%
N-P	VGG19+LSTM	100%	97.06%	98.51%
N-S	VGG16+MLP	100%	97.62%	98.80%
N-V	VGG19+LSTM	100%	100%	100%
D-P	VGG19+LSTM	76.19%	94.12%	84.21%
D-S	VGG16+MLP	83.67%	97.62%	90.11%
D-V	VGG16+MLP	92.42%	98.39%	95.31%
	VGG16+LSTM	92.42%	98.39%	95.31%
P-S	VGG16+MLP	86.96%	95.24%	90.91%
P-V	VGG19+MLP	88.41%	98.39%	93.13%
S-V	VGG19+MLP	95.38%	100%	97.64%

6.5.3 Experiment on Medetec Dataset

A classification between all the classes is performed on the Medetec dataset. Table 6.11 shows the results of this three-wound-class classification (D vs. P vs. A+V). We achieve the highest accuracy of 86.67% with the multi-modal (WMC) network using the VGG19+MLP and VGG19+LSTM combinations, where the highest accuracy achieved from WLC and WIC networks is 85.56% and 82.22% using both MLP and LSTM, and VGG16 networks, respectively.

Table 6.11: Results of three-wound-class classifications (D vs. P vs. A+V) on the Medetec Dataset.

Input	Model	Accuracy
Location	MLP	85.56%
	LSTM	85.56%
Image	VGG16	82.22%
	VGG19	77.78%
Image + Location	VGG16+MLP	85.56%
	VGG19+MLP	**86.67%**
	VGG16+LSTM	85.56%
	VGG19+LSTM	**86.67%**

6.5.4 Experiment on AZHMT Dataset

A classification between all the classes is performed on the AZHMT dataset. Table 6.12 shows the results of this six-class classification (BG vs. N vs. D vs. P vs. S vs. A+V). We achieve the highest accuracy of 83.04% with the multi-modal (WMC) network using the VGG19+LSTM combination. The highest accuracy of WLC and WIC networks is 71.30% and 72.22% using LSTM and VGG19 networks, respectively.

Finally, a four-wound-class classification is performed on the AZHMT dataset. The classification is done among D, P, S, and A+V classes. We achieved the highest accuracy of 84.31% with the multi-modal (WMC) network using the VGG19+MLP combination. The highest accuracy achieved from WLC and WIC networks is 78.83% and 68.61% using LSTM and VGG16 networks. Table 6.13 shows the detailed results of this four-wound-class classification.

Table 6.12: Six-class classification (BG vs. N vs. D vs. P vs. S vs. A+V) results on the AZHMT Dataset.

Input	Model	Accuracy
Location	MLP	69.44%
	LSTM	71.30%
Image	VGG16	67.59%
	VGG19	72.22%
Image + Location	VGG16+MLP	81.17%
	VGG19+MLP	81.79%
	VGG16+LSTM	72.22%
	VGG19+LSTM	**83.04%**

Table 6.13: Four-wound-class classification (D vs. P vs. S vs. A+V) results on the AZHMT Dataset.

Input	Model	Accuracy
Location	MLP	78.10%
	LSTM	78.83%
Image	VGG16	68.61%
	VGG19	63.14%
Image + Location	VGG16+MLP	79.56%
	VGG19+MLP	**84.31%**
	VGG16+LSTM	68.25%
	VGG19+LSTM	68.98%

6.5.5 Result Comparison with Previous Works

Classification results depend on many factors like dataset, model, training-validation-testing split, balanced or unbalanced dataset, resources used for training, etc. Though the datasets and other factors between our work and previous classification works are not the same, this section mainly focuses on how the multimodality using both image and location data can improve the classification accuracy. The comparison with the previous works is only made if all the classes of that work's dataset are present in our dataset. Rostami et al.'s work [79]'s dataset is most like the work presented in this chapter. Alongside [79], the classifications performed in [58], [78], and [60] have the classes that are present in our dataset. A detailed comparison between previous works and our current work is shown in Table 6.14.

The reasons why other related works are not considered in this comparison are: [70] performs burn vs. pressure ulcer classification, and our datasets do not contain any burn images; [63] performs binary classification of ischemia vs. non-ischemia and infection vs. non-infection on DFU images, which is not compatible with our datasets; [59] performs binary classifications between such kind of wounds (wound, infection (SSI), granulation tissue, etc.), which are not present in our datasets; and [80] performs multi-class wound classifications among diabetic, lymphovascular, pressure injury, and surgical wounds and our datasets do not contain the lymphovascular wound type.

Table 6.14: Comparison among the previous works and the present work on wound type classifications.

Work	Classification	Evaluation Metrics	Previous Work			Present Work		
			Model	Dataset	Result	Model	Dataset	Result
Goyal et al. [58]	Healthy Skin Vs. DFU Skin (N vs. D)	Accuracy	DFUNet	A dataset containing 397 wound images	92.5%	VGG19 +LSTM	AZH	**100%**
Aguirre et al. [78]	VLU versus non-VLU (N vs. V, D vs. V, P vs. V, S vs. V)	Accuracy	VGG19	A dataset of 300 wound images	85%	**N-V:** VGG19 +LSTM	AZH	**100%**
						D-V: VGG16 +MLP & VGG16 +LSTM		**94.44%**
						P-V: VGG19 +MLP		**90.63%**
						S-V: VGG19 +MLP		**97.12%**
Alzubaidi et al. [60]	Normal Skin Vs. Abnormal (DFU) Skin (N vs. D)	F1-Score	DFU_QUTNet + SVM	A dataset containing 754-foot images	94.5%	VGG19 +LSTM	AZH	**100%**
Rostami et al. [79]	S-V	Accuracy	An end-to-end Ensemble DCNN-based Classifier	A new dataset containing 538 wound images	96.4%	VGG19 +MLP	AZH	**97.12%**
	D-S-V				91.9%	VGG19+MLP		**92.00%**
	BG-N-D-V				89.41%	VGG19+MLP		**95.57%**
	BG-N-P-V				86.57%	VGG16 +LSTM		**92.47%**
	BG-N-S-V				92.20%	VGG16+MLP		**94.16%**
	BG-N-D-P				80.29%	VGG19 +LSTM		**89.23%**

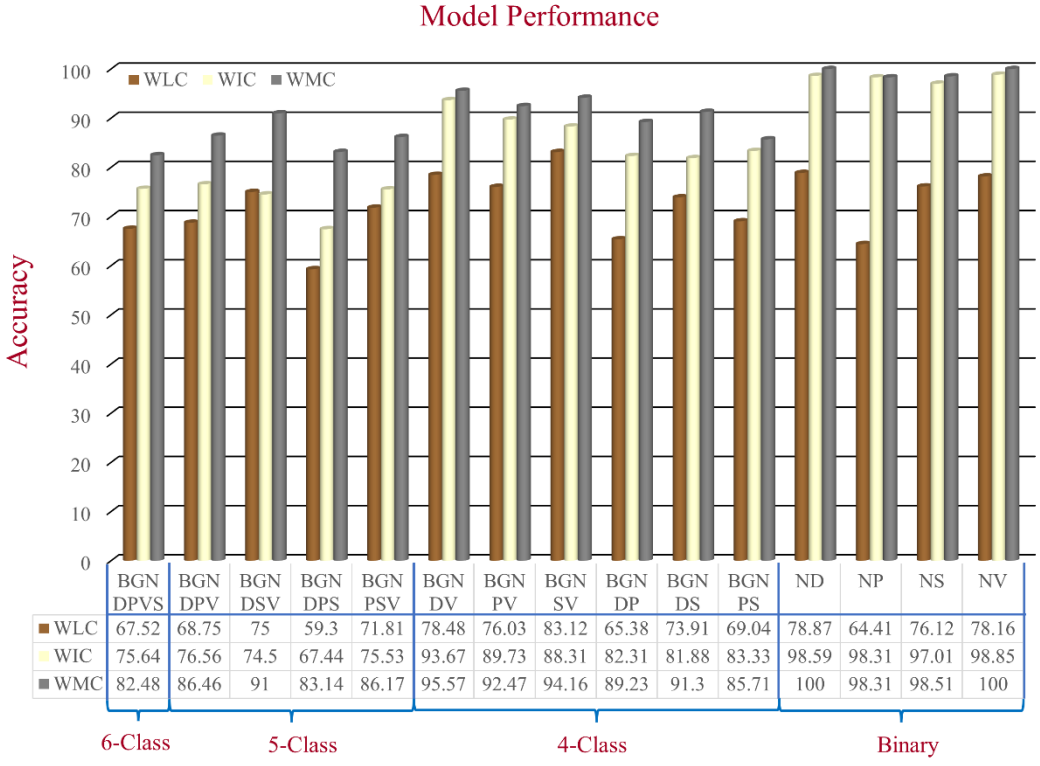
Work	Classification	Evaluation Metrics	Previous Work			Present Work		
			Model	Dataset	Result	Model	Dataset	Result
	BG-N-D-S				90.98%	VGG19+MLP		**91.30%**
	BG-N-P-S				84.12%	VGG16+MLP		**85.71%**
	BG-N-D-P-V				79.76%	VGG16+MLP		**84.46%**
	BG-N-D-S-V				84.94%	VGG16 +LSTM		**91.00%**
	BG-N-D-P-S				81.49%	VGG16+MLP		**83.14%**
	BG-N-P-S-V				83.53%	VGG19+MLP		**86.17%**
	BG-N-D-P-S-V				68.69%	VGG19+MLP		**82.48%**

6.6 Discussion

In all the experiments performed in this chapter, there are two types of classifications: 1) mixed-class classifications (e.g., three-class classification, five-class classification, etc.), and 2) wound-class classifications (e.g., four wound-class classification, three wound-class classifications, etc.). The wound-class classification does not contain any non-wound classes (i.e., normal skin and background), and they are more challenging to classify than the mixed-class classification. This section will discuss the classifications performances, comparison with state-of-the-art results, limitations, and how to overcome them.

6.6.1 Performance Analysis and the Power of Multimodality

On the AZH dataset, for mixed-class classifications, we performed one six-class, four five-class, six four-class, and four binary classifications. We performed one four-wound-class, four three-wound-class, and six binary classifications for wound-class classifications. From Table 6.5 to Table 6.9, the same consistency of the model performances is observed, where the best to worst results are achieved by WMC, WIC, and WLC classifiers, respectively. Though a single model of WLC or WIC or a single combination of WMC does not always produce the best performance, the WMC classifier always performs the best compared to the WIC or WLC classifiers. The same pattern can also be seen in the wound class classifications.



	BGN DPVS	BGN DPV	BGN DSV	BGN DPS	BGN PSV	BGN DV	BGN PV	BGN SV	BGN DP	BGN DS	BGN PS	ND	NP	NS	NV
WLC	67.52	68.75	75	59.3	71.81	78.48	76.03	83.12	65.38	73.91	69.04	78.87	64.41	76.12	78.16
WIC	75.64	76.56	74.5	67.44	75.53	93.67	89.73	88.31	82.31	81.88	83.33	98.59	98.31	97.01	98.85
WMC	82.48	86.46	91	83.14	86.17	95.57	92.47	94.16	89.23	91.3	85.71	100	98.31	98.51	100

Figure 6.6: Performance comparison of mixed-class classification among the best models from each category (location -WLC, image-WIC, and multimodality-WMC) on the AZH Dataset.

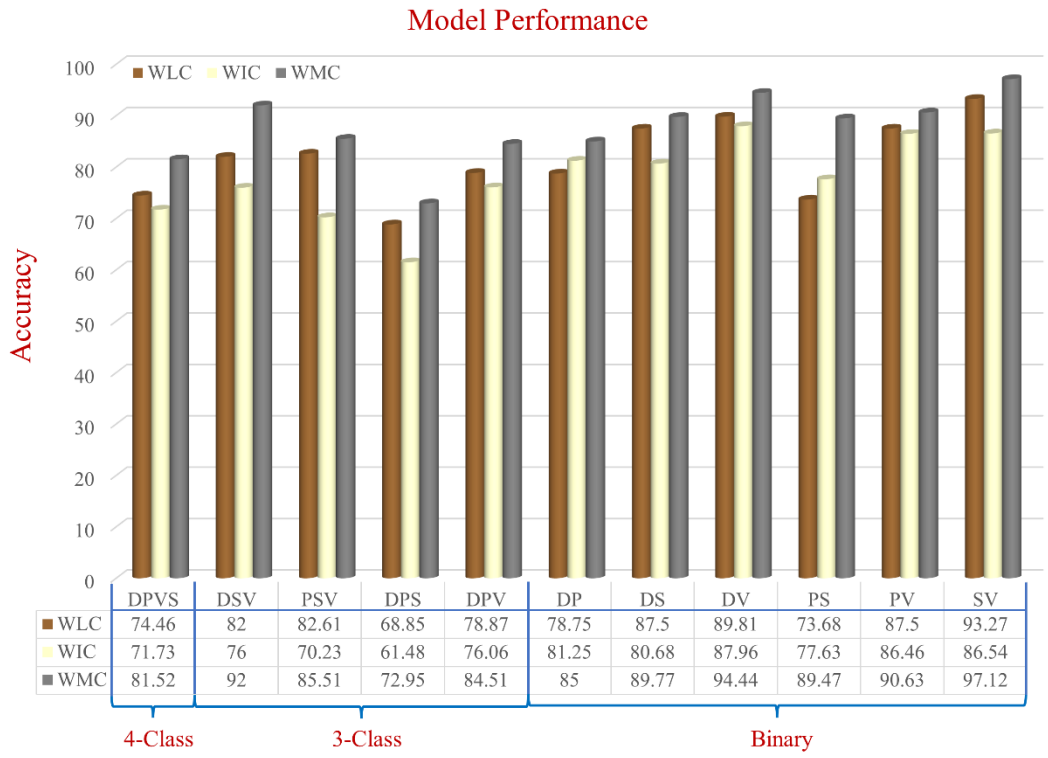


Figure 6.7: Performance comparison of wound-class classification among the best models from each category (location -WLC, image-WIC, and multimodality-WMC) on the AZH Dataset.

The performance comparison of mixed-class classifications among the best models from each category (location, image, and multimodality) is shown in Figure 6.6. The performance comparison among the best models of wound-class classifications from each category (location, image, and multimodality) is shown in Figure 6.7. From Figure 6.6, the lowest accuracy is produced by BGNDPS (83.14%), and from Figure 6.7, the most insufficient accuracy is produced by DPS (72.95%). So, separating diabetic, pressure, and surgical wound is the hardest, according to our experiments. Also, from Figure 6.7, among all binary classifications, D vs. P has the lowest accuracy of 85%. So, we can say that differentiation between diabetic and pressure wounds is the most complicated task. From Figure 6.6, the highest accuracy is achieved by ND, NP, NS, and NV classifications with 100%, 98.31%, 98.51%, and 100%, respectively. Also, from Figure 6.7, the

highest accuracy is achieved by SV classification with 97.12% accuracy. So, differentiating between normal skin and other wound types (D, V, S, and P) and differentiating between surgical wounds and venous leg ulcers are the most straightforward classifications task for our developed WMC classifier. Finally, from Figures 6.6 and 6.7, we can see that multimodality using wound image and location (WMC) performs best in comparison with single (image or location) modality (WLC or WIC) in all scenarios on the AZH dataset; and mixed-class classification results are comparatively higher than wound-class classification results.

6.6.2 Robustness Testing

To evaluate the robustness of our developed WMC classifier, we experimented on a publicly available dataset named Medetec Dataset, which has an entirely different data collection and distribution than our collected and developed AZH Dataset. We perform only one wound-type classification among all three classes (D, P, and A+V) on this dataset. As a result, the highest accuracies achieve by WLC, WIC, and WMC classifiers are 85.56%, 82.22%, and 86.67%, respectively. So, the highest accuracy is achieved by the WMC classifier, which proves that the WMC works well on different datasets with separate distributions.

6.6.3 The Effect of Bigger Dataset

We developed a mixed and more extensive dataset named AZHMT to test the effect of adding more data points to our model performance. AZHMT is a mixed dataset containing wound image and location data from AZH and Medetec datasets. On the AZHMT dataset, we perform one six-mixed-class classification (BG-N-D-P-S-A+V) and one four-wound-class classification (D-P-S-A+V). Comparing these results of AZH and AZHMT datasets, we see that with the AZHMT dataset, we achieved higher accuracy than the AZH dataset. A comparison between the highest results (accuracy) of AZH and AZHMT datasets is shown in Figure 6.8. The results are from the multi-modal network (WMC), as it outperforms all the single modal (WIC and WLC) networks. The AZHMT dataset has 0.56% more accuracy for the six-class classification than the AZH dataset. The AZHMT dataset has 2.79% more accuracy for the four-wound-class classification than the AZH dataset. Here, AZHMT contains more data than the AZH dataset, which is an advantage for training deep learning models; but AZHMT also contains mixed data from two sources, which makes the dataset more challenging to classify; AZHMT also contains a mixed data on a single class (arterial and venous ulcer combination), which may also impact the results. Regardless of having some disadvantages of the mixed dataset, this comparison proves that increasing data points improve model performance.

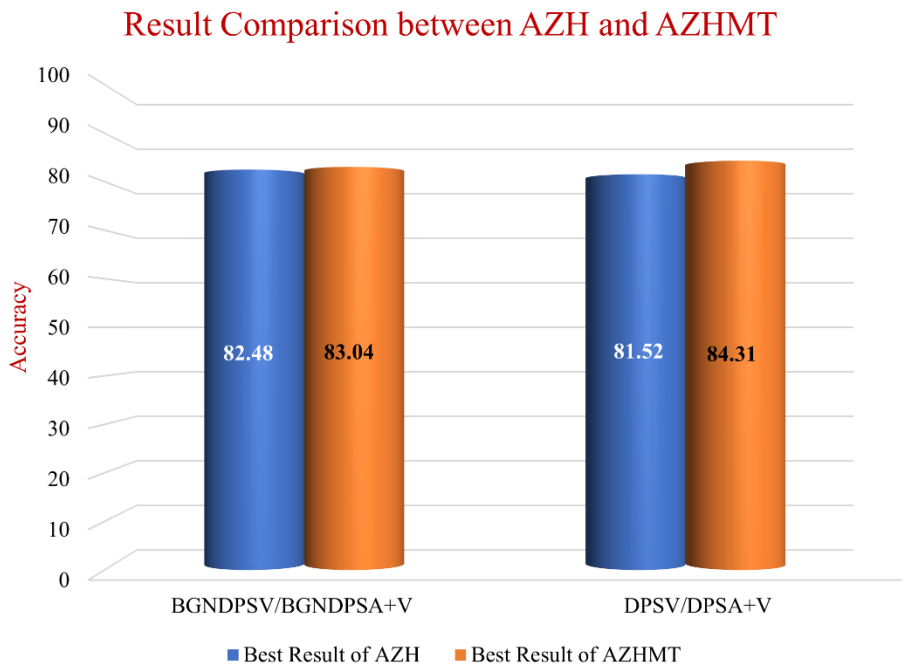


Figure 6.8: Comparison between the highest results (accuracy) of the AZH and the AZHMT datasets.

6.6.4 Comparison with Previous Works

From Table 6.14, we can see that our work outperforms all the previous works by a good margin. As mentioned earlier, this comparison is not perfect as factors like dataset, model, training-validation-testing split, balance ness of the dataset, resources used for training, etc., are not the same as the previous works. But this comparison proves that multimodality using wound image and location can improve the wound classification results. We achieved a 7.5% improvement in accuracy for classifying Healthy Skin Vs. DFU Skin (N Vs. D) from Goyal et al.'s work [58] on our AZH dataset. Compared to Aguirre et al.'s work [78] of classifying VLU versus non-VLU (V vs. [N or D or P or S]) wounds, we achieved a significant 5.63% to 15% improvement in accuracy with the AZH dataset. In this experiment, we got an improvement of 5.63% for VLU vs. PU, 9.44%

for VLU vs. DFU, 12.12% for VLU vs. Surgical, and 15% for VLU vs. Normal skin. Our developed classifier outperformed Alzubaidi et al.'s work [60] on Normal Skin Vs. Abnormal (DFU) Skin (N vs. D) classification with 5.5% improvement in F1-score for the AZH experiment. Finally, compared to our previous work [79], there are 13 similar experiments in our present work. We achieved a significant improvement with the multi-modal WMC network in all these experiments. In these 13 experiments, the accuracy improvement using WMC classifier from our previous work are: 1) 0.72% improvement in SV classification, 2) 0.1% improvement in DSV classification, 3) 6.16% improvement in BGNDV classification, 4) 5.9% improvement in BGNPV classification, 5) 1.96% improvement in BGNSV classification, 6) 8.94% improvement in BGNDP classification, 7) 0.32% improvement in BGNDS classification, 8) 1.59% improvement in BGNPS classification, 9) 4.7% improvement in BGNDPV classification, 10) 6.06% improvement in BGNDSV classification, 11) 1.65% improvement in BGNDPS classification, 12) 2.64% improvement in BGNPSV classification, and 13) 13.79% improvement in BGNDPSV classification. Both of these works have some pros and cons: in our previous work, we have a balanced dataset (all classes had the same no of images), where the current work has an unbalanced dataset (Table 6.2); the previous work uses a very sophisticated ensemble classifier for image classification, where this work uses simple transfer learning with available DNN networks (VGG16, VGG19, etc.); the previous work only used wound images for training the classifier, where the current network uses both wound images and their corresponding locations for developing the classifier. Overall, this work outperforms all the previous works by a good difference.

6.6.5 Limitations and Scope of Improvement

In Figure 6.6, the WLC network's performance is inferior compared to the WIC and WMC network. One important reason is that there are some overlaps among the normal (healthy) skin and other wound classes, as the normal skin is cropped from the wound images. In one patient's wound image, a non-infected (normal) skin can be infected in another patient's wound image, which produces these overlaps and thus decreases the WLC performance. Figure 6.7 shows that the WLC network's performance is better than the WIC network as there is no normal skin (N) class in these classifications. The WLC network performance can be improved by increasing the number of data points, which can help increase the WMC network's performance in the long run. Figure 6.9 shows some examples of location overlapping among different classes.

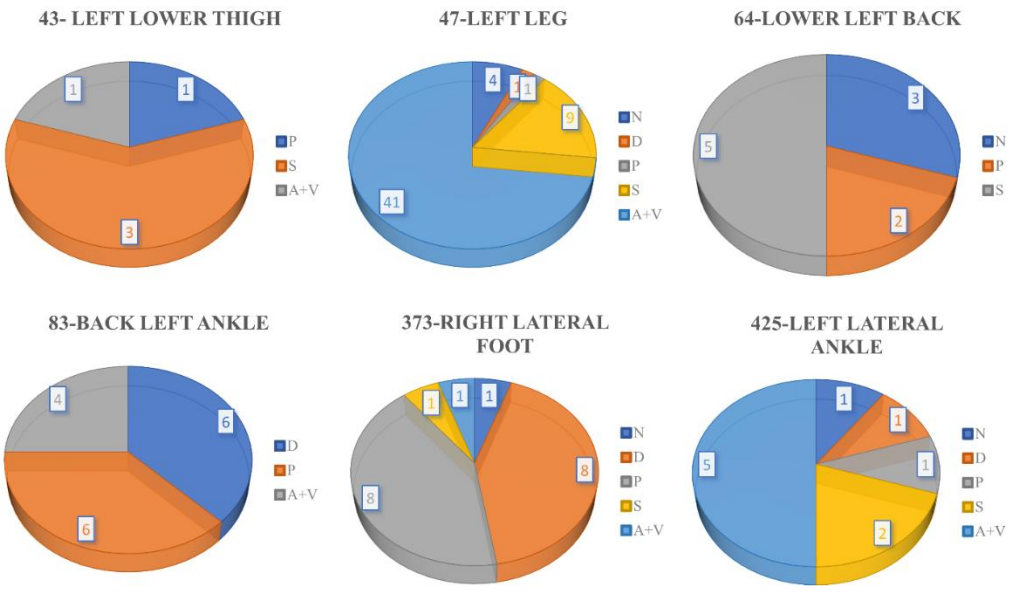


Figure 6.9: Examples of location overlap on the AZHMT dataset.

6.7 Conclusion

This chapter developed a multi-modal wound classifier (WMC) network using wound images and their corresponding locations to classify wounds into different classes. To the best of our knowledge, it is the first developed multi-modal network that uses images and locations for wound classification. This research is also the first work that classifies wounds according to their locations. We also developed a body map to help clinicians document the wound locations in the patient's record to prepare the location data. The developed body map is currently used in the AZH wound center for location tagging to avoid inconsistency with location information. Three datasets with wound images and their corresponding locations are also developed and labeled by wound specialists of AZH wound center to perform many wound classification experiments. To the best of our knowledge, many experiments with a range of binary to six-class classifications are conducted in three datasets, where many wound classifications are never performed before. The results produced by the WMC network are much better than the results produced from the WIC or WLC networks, and these results beat all the previous experimental results.

Many wound patients lack access to specialized wound care and updated instructions due to a scarcity of well-trained wound experts in primary and rural healthcare settings. Remote telemedicine system advancements may considerably assist patients in remote regions, particularly in rural areas, by providing improved diagnostic guidance, which is especially important in pandemics like COVID-19 [23]. An intelligent system can help with wound care in various ways, including enhanced accuracy, decreased effort and economic burden, standardized diagnosis and treatment, and improved patient care quality. [24]. Keeping this importance of an intelligent wound care system, we are building an automated wound management tool that can store patient information and predict wound healing probability from the given information. The current

multimodality work is the basic block of this ultimate wound healing prediction system. Figure 6.10 shows an idea of a future wound prognosis system that uses multimodality deep learning as the core model to predict wound healing time. The input will come from wound image, body map (corresponding location), and patient's electronic health record (EHR). The modalities will be images, categorical data, and continuous data. Different features will be the input of the multi-modal network from various sources. Some example features for this model: from images, we will get color, texture, wound area, wound type, wound severity, wound tissue type; from body map, we will get corresponding wound location; and from EHR we will get patient's age, sex, blood pressure, etc. Some features may be obtained by passing the input through our developed deep learning models (i.e., localizer, segmenter, classifier, etc.). Overall, multi-modal deep learning model is the key to build an AI driven automated system that can perform wound healing prediction.

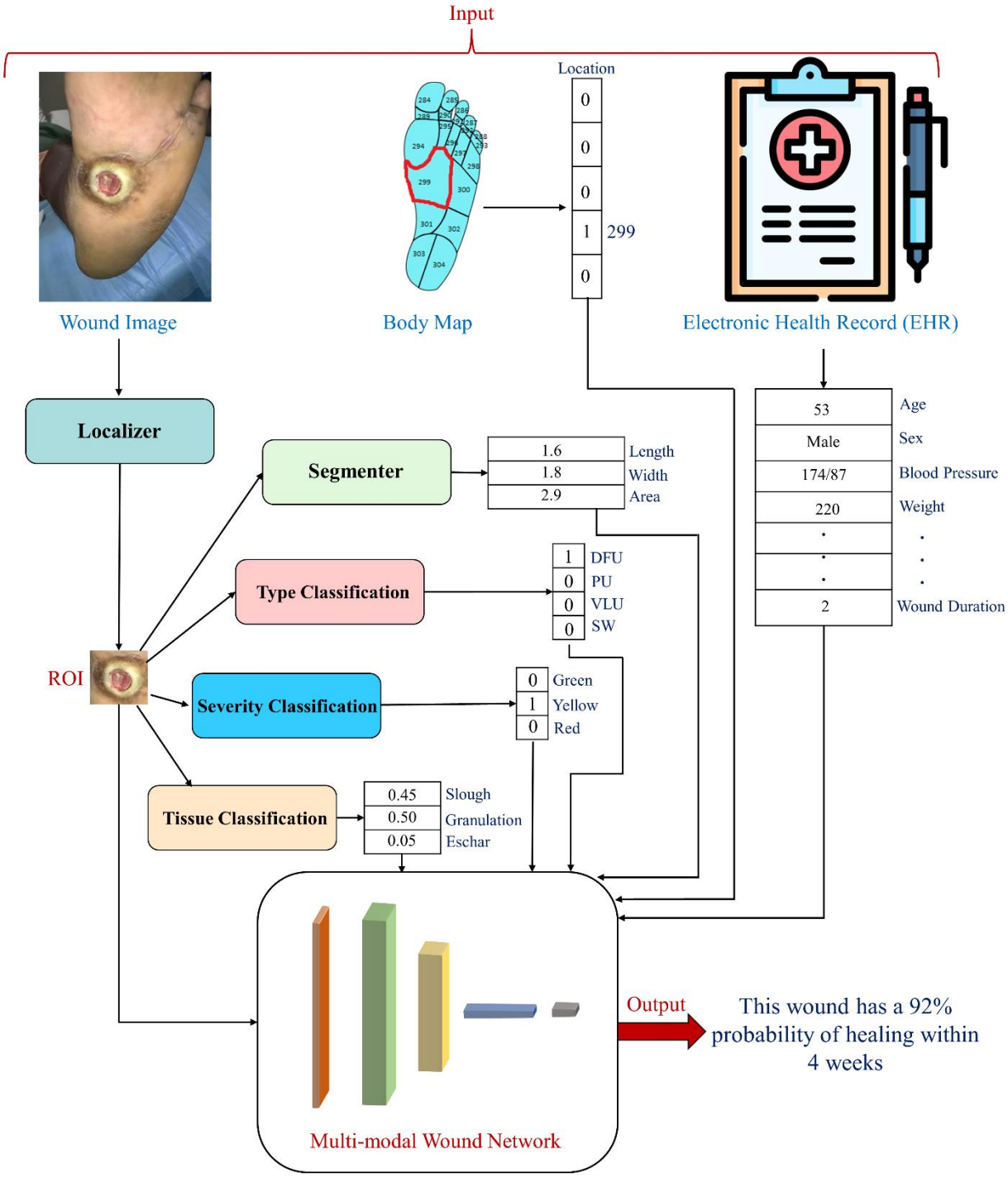


Figure 6.10: The future implication of the developed multi-modality network to predict wound healing time.

Chapter 7

Conclusion

7.1 Conclusion

This thesis performs several medical image classifications using deep neural networks. A total of four classifications, including osteosarcoma classification, burn wound classification, wound severity classification, and wound type classification, are performed on medical images. In addition, a significant pre-processing step of wound image processing named the wound region of interest (ROI) detector is also automated using a deep learning model.

In osteosarcoma classification, histological images are classified into viable tumors, necrotic tumors, and non-tumor. Transfer learning models with adjustments to adapt the input data are used to develop the network. The developed network shows a better performance than the previous works.

Wound ROIs are detected using deep learning models like YOLOv3 and SSD. ROI detection is a necessary step for wound image segmentation and classification works. The raw images collected from the clinic contain a lot of unnecessary backgrounds, which will not only hamper the classification or segmentation performance but also raise an issue of data privacy as medical images are very sophisticated to share publicly. Our developed ROI detector helps to prevent these issues and significantly improve classification performance. This work is necessary to the literature as most of the previous wound segmentation and classification works use manually cropped ROIs, which is not efficient in terms of both time and cost.

Burn wounds are classified into binary classes (graft versus non-graft) and multiclass (full-thickness, deep dermal, and superficial dermal) using deep learning models. Transfer learning and end-to-end learning models are used to perform these classification tasks. As a result, better results from the previous works are achieved.

Wound severity is detected by using deep learning models. The stage of importance, including three classes, green, yellow, and red, are determined where green represents low severity cases and red represents most severe cases. We used transfer learning and stacked models to perform this classification task. This work can help wound professionals in treatment plan making. One multiclass and three binary classifications are performed on this developed dataset to differentiate among wound severity levels. Unfortunately, excellent classification performance is not achieved due to the scarcity of data and noise level present in the dataset.

Finally, a novel deep learning model is presented for wound type classification using both wound images and their corresponding locations. A multimodal (WMC) network is developed as the dataset contains both image and categorical (wound location) data. This developed classifier combines the power of image classification with CNN models like VGG16, VGG19, AlexNet, etc., and location classification with MLP and LSTM networks. This is a unique contribution to literature as this is the first work to perform multimodality with wound images and their corresponding locations. This developed wound classifier shows a significant improvement from existing image-based wound type classifiers.

7.2 Future Directions

The future directions for this thesis are:

- Refining the burn wound classifier performance by developing a more extensive and improved dataset. As collecting medical data is a considerable challenge, GAN models can be used to create synthetic burn wound images in the future.
- Improving the wound severity classification performance by adapting the network for the dataset's needs. This includes involving more sophisticated features like wound size, wound number, wound depth, etc. A better and bigger dataset is required to perform this task, including expert's (physician or nurse) annotation.
- There are some overlaps in the wound location data, for which the WLC network produces lower accuracy than WIC and WMC networks. Therefore, increasing the number of data can improve the location (WLC) classifier's performance.
- Using the concept of multi-modality, an AI-driven automated wound analysis system can be developed, which will take input from different modalities like images, electronic health records, etc.

Bibliography

[1] F. Zhao, Y. Chen, Y. Hou, and X. He, “Segmentation of blood vessels using rule-based and machine-learning-based methods: a review,” *Multimed. Syst.*, vol. 25, no. 2, pp. 109–118, Dec. 2017, doi: 10.1007/S00530-017-0580-7.

[2] J. Duryea, J. Li, C. G. Peterfy, C. Gordon, and H. K. Genant, “Trainable rule-based algorithm for the measurement of joint space width in digital radiographic images of the knee,” *Med. Phys.*, vol. 27, no. 3, pp. 580–591, Mar. 2000, doi: 10.1118/1.598897.

[3] A. Singh, Narina Thakur, and Aakanksha Sharma, “A review of supervised machine learning algorithms ,” in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2016, pp. 1310–1315, Accessed: Aug. 11, 2021. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7724478.

[4] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, “Supervised machine learning: A review of classification techniques,” *Emerg. Artif. Intell. Appl. Comput. Eng.*, vol. 160, pp. 3–24, 2007.

[5] A. Olaode, G. Naghdy, and C. Todd, “Unsupervised classification of images: A review,” *Int. J. Image Process.*, vol. 8, no. 5, pp. 325–342, 2014.

[6] M. L. Giger, “Machine Learning in Medical Imaging,” *J. Am. Coll. Radiol.*, vol. 15, no. 3, pp. 512–520, Mar. 2018, doi: 10.1016/J.JACR.2017.12.028.

[7] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, “Deep Learning for Computer Vision: A Brief Review,” *Comput. Intell. Neurosci.*, vol. 2018, 2018, doi: 10.1155/2018/7068349.

[8] G. Litjens *et al.*, “A survey on deep learning in medical image analysis,” *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/J.MEDIA.2017.07.005.

[9] M. Bakator and D. Radosav, “Deep Learning and Medical Diagnosis: A Review of Literature,” *Multimodal Technol. Interact. 2018, Vol. 2, Page 47*, vol. 2, no. 3, p. 47, Aug. 2018, doi: 10.3390/MTI2030047.

[10] A. J. Chou, D. S. Geller, and R. Gorlick, “Therapy for Osteosarcoma,” *Pediatr. Drugs 2008 105*, vol. 10, no. 5, pp. 315–327, Aug. 2012, doi: 10.2165/00148581-200810050-00005.

[11] C. A. S. Arndt and W. M. Crist, “Common Musculoskeletal Tumors of Childhood and Adolescence,” *N. Engl. J. Med.*, vol. 341, no. 5, pp. 342–352, Oct. 2008, doi: 10.1056/NEJM199907293410507.

[12] P. P. Lin and S. Patel, “Osteosarcoma,” *Bone Sarcoma*, pp. 75–97, Jan. 2013, doi: 10.1007/978-1-4614-5194-5_5.

[13] D. S. Geller and R. Gorlick, “Osteosarcoma: A Review of Diagnosis, Management, and Treatment Strategies,” *Clin. Adv. Hematol. Oncol.*, vol. 8, no. 10, 2010.

[14] N. Wahab, A. Khan, and Y. S. Lee, “Transfer learning based deep CNN for segmentation and detection of mitoses in breast cancer histopathological images,” *Microscopy*, vol. 68, no. 3, pp. 216–233, Jun. 2019, doi: 10.1093/JMICRO/DFZ002.

[15] P. Picci, “Osteosarcoma (Osteogenic sarcoma),” *Orphanet J. Rare Dis. 2007 21*, vol. 2, no. 1, pp. 1–4, Jan. 2007, doi: 10.1186/1750-1172-2-6.

[16] G. Litjens *et al.*, “Deep learning as a tool for increased accuracy and efficiency of

histopathological diagnosis," *Sci. Reports 2016 61*, vol. 6, no. 1, pp. 1–11, May 2016, doi: 10.1038/srep26286.

[17] K. Smith and T. Nolan, "Osteosarcoma data from UT Southwestern/UT Dallas for Viable and Necrotic Tumor Assessment (Osteosarcoma-Tumor-Assessment)," 2019. https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=52756935.

[18] C. K. Sen, "Human Wound and Its Burden: Updated 2020 Compendium of Estimates," *Adv. Wound Care*, vol. 10, no. 5, pp. 281–292, Mar. 2021, doi: 10.1089/WOUND.2021.0026.

[19] "Diabetic Foot: Facts & Figures." https://diabeticfootonline.com/diabetic-foot-facts-and-figures/ (accessed Jun. 02, 2021).

[20] E. A. Nelson and U. Adderley, "Venous leg ulcers," *BMJ Clin. Evid.*, vol. 2016, no. March 2014, pp. 1–36, 2016.

[21] "Preventing Pressure Ulcers in Hospitals." https://www.ahrq.gov/patient-safety/settings/hospital/resource/pressureulcer/tool/pu1.html (accessed Jun. 04, 2021).

[22] B. M. Gillespie *et al.*, "Setting the surgical wound care agenda across two healthcare districts: A priority setting approach," *Collegian*, vol. 27, no. 5, pp. 529–534, 2020, doi: 10.1016/j.colegn.2020.02.011.

[23] F. L. Bowling *et al.*, "Remote assessment of diabetic foot ulcers using a novel wound imaging system," *Wound Repair Regen.*, vol. 19, no. 1, pp. 25–30, Jan. 2011, doi: 10.1111/J.1524-475X.2010.00645.X.

[24] E. Pasero and C. Castagneri, "Application of an automatic ulcer segmentation algorithm,"

in *RTSI 2017 - IEEE 3rd International Forum on Research and Technologies for Society and Industry, Conference Proceedings*, Oct. 2017, pp. 1–4, doi: 10.1109/RTSI.2017.8065954.

[25] K. H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nat. Biomed. Eng.*, vol. 2, no. 10, pp. 719–731, 2018, doi: 10.1038/s41551-018-0305-z.

[26] F. ROSENBLATT, "PRINCIPLES OF NEURODYNAMICS. PERCEPTRONS AND THE THEORY OF BRAIN MECHANISMS," 1961, Accessed: Mar. 29, 2022. [Online]. Available: https://apps.dtic.mil/sti/citations/AD0256582.

[27] Sepp Hochreiter and Jurgen Schmidhuber, "LONG SHORT-TERM MEMORY," *Neural Comput.* , pp. 1735–1780, 1997, Accessed: Mar. 29, 2022. [Online]. Available: http://www.bioinf.jku.at/publications/older/2604.pdf.

[28] Krizhevsky, Alex, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems 25*, 2012, pp. 1097–1105.

[29] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, Sep. 2014, Accessed: Jul. 16, 2021. [Online]. Available: https://arxiv.org/abs/1409.1556v6.

[30] C. Szegedy *et al.*, "Going Deeper with Convolutions," in *IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, Accessed: Jul. 16, 2021. [Online]. Available: http://image-net.org/challenges/LSVRC/2015/.

[32] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger, "Densely Connected Convolutional Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708, Accessed: Mar. 29, 2022. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.html.

[33] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," Accessed: Mar. 29, 2022. [Online]. Available: www.aaai.org.

[34] Francois Chollet, "Xception: Deep Learning With Depthwise Separable Convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1251–1258, Accessed: Mar. 29, 2022. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Chollet_Xception_Deep_Learning_CVPR_2017_paper.html.

[35] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning Transferable Architectures for Scalable Image Recognition," in *IEEE conference on computer vision and pattern recognition*, 2018, pp. 8697–8710.

[36] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," 2018.

[37] W. Liu *et al.*, “SSD: Single Shot MultiBox Detector,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9905 LNCS, pp. 21–37, 2016, doi: 10.1007/978-3-319-46448-0_2.

[38] J. Redmon and A. Farhadi, “YOLOv3: An Incremental Improvement,” *arXiv*, Apr. 2018, Accessed: Oct. 13, 2021. [Online]. Available: https://arxiv.org/abs/1804.02767v1.

[39] Vaibhav Jayaswal, “Performance Metrics: Confusion matrix, Precision, Recall, and F1 Score | by Vaibhav Jayaswal | Towards Data Science.” https://towardsdatascience.com/performance-metrics-confusion-matrix-precision-recall-and-f1-score-a8fe076a2262 (accessed Mar. 30, 2022).

[40] Jonathan Hui, “mAP (mean Average Precision) for Object Detection | by Jonathan Hui | Medium.” https://jonathan-hui.medium.com/map-mean-average-precision-for-object-detection-45c121a31173 (accessed Mar. 30, 2022).

[41] Jakub Czakon, “F1 Score vs ROC AUC vs Accuracy vs PR AUC: Which Evaluation Metric Should You Choose? - neptune.ai.” https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc (accessed Mar. 30, 2022).

[42] J. De Matos, A. D. S. Britto, L. E. S. Oliveira, and A. L. Koerich, “Double Transfer Learning for Breast Cancer Histopathologic Image Classification,” *Proc. Int. Jt. Conf. Neural Networks*, vol. 2019-July, Jul. 2019, doi: 10.1109/IJCNN.2019.8852092.

[43] N. Wahab, A. Khan, and Y. S. Lee, “Transfer learning based deep CNN for segmentation and detection of mitoses in breast cancer histopathological images,” *Microscopy*, vol. 68, no. 3, pp. 216–233, Jun. 2019, doi: 10.1093/JMICRO/DFZ002.

[44] M. Yanagawa *et al.*, “Application of deep learning (3-dimensional convolutional neural network) for the prediction of pathological invasiveness in lung adenocarcinoma: A preliminary study,” *Medicine (Baltimore).*, vol. 98, no. 25, p. e16119, Jun. 2019, doi: 10.1097/MD.0000000000016119.

[45] C. Sun, A. Xu, D. Liu, Z. Xiong, F. Zhao, and W. Ding, “Deep Learning-Based Classification of Liver Cancer Histopathology Images Using Only Global Labels,” *IEEE J. Biomed. Heal. Informatics*, vol. 24, no. 6, pp. 1643–1651, Jun. 2020, doi: 10.1109/JBHI.2019.2949837.

[46] Y. Celik, M. Talo, O. Yildirim, M. Karabatak, and U. R. Acharya, “Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images,” *Pattern Recognit. Lett.*, vol. 133, pp. 232–239, May 2020, doi: 10.1016/J.PATREC.2020.03.011.

[47] K. M. Hosny, M. A. Kassem, and M. M. Foaud, “Skin Cancer Classification using Deep Learning and Transfer Learning,” *2018 9th Cairo Int. Biomed. Eng. Conf. CIBEC 2018 - Proc.*, pp. 90–93, Feb. 2019, doi: 10.1109/CIBEC.2018.8641762.

[48] L. Huang, W. Xia, B. Zhang, B. Qiu, and X. Gao, “MSFCN-multiple supervised fully convolutional networks for the osteosarcoma segmentation of CT images,” *Comput. Methods Programs Biomed.*, vol. 143, pp. 67–74, May 2017, doi: 10.1016/J.CMPB.2017.02.013.

[49] R. Zhang, L. Huang, W. Xia, B. Zhang, B. Qiu, and X. Gao, “Multiple supervised residual network for osteosarcoma segmentation in CT images,” *Comput. Med. Imaging Graph.*, vol. 63, pp. 1–8, Jan. 2018, doi: 10.1016/J.COMPMEDIMAG.2018.01.006.

[50] Mishra Rashika, Daescu Ovidiu, Leavey Patrick, Rakheja Dinesh, and Sengupta Anita, "Convolutional Neural Network for Histopathological Analysis of Osteosarcoma," *J. Comput. Biol.*, vol. 25, no. 3, pp. 313–325, Mar. 2018, doi: 10.1089/CMB.2017.0153.

[51] H. B. Arunachalam *et al.*, "Viable and necrotic tumor assessment from whole slide images of osteosarcoma using machine-learning and deep-learning models," *PLoS One*, vol. 14, no. 4, p. e0210706, Apr. 2019, doi: 10.1371/JOURNAL.PONE.0210706.

[52] F. Chollet, "Image data preprocessing," *Keras*. https://keras.io/api/preprocessing/image/ (accessed Jul. 12, 2020).

[53] "Keras Applications." https://keras.io/api/applications/ (accessed Jul. 16, 2021).

[54] E. S. Papazoglou, L. Zubkov, X. Mao, M. Neidrauer, N. Rannou, and M. S. Weingarten, "Image analysis of chronic wounds for determining the surface area," *Wound Repair Regen.*, vol. 18, no. 4, pp. 349–358, Jul. 2010, doi: 10.1111/J.1524-475X.2010.00594.X.

[55] N. D. J. Hettiarachchi, R. B. H. Mahindaratne, G. D. C. Mendis, H. T. Nanayakkara, and N. D. Nanayakkara, "Mobile based wound measurement," *IEEE EMBS Spec. Top. Conf. Point-of-Care Healthc. Technol. Synerg. Towar. Better Glob. Heal. PHT 2013*, pp. 298–301, 2013, doi: 10.1109/PHT.2013.6461344.

[56] M. C. Chang *et al.*, "Multimodal Sensor System for Pressure Ulcer Wound Assessment and Care," *IEEE Trans. Ind. Informatics*, vol. 14, no. 3, pp. 1186–1196, Mar. 2018, doi: 10.1109/TII.2017.2782213.

[57] K. Wantanajittikul, S. Auephanwiriyakul, N. Theera-Umpon, and T. Koanantakool, "Automatic segmentation and degree identification in burn color images," in *BMEiCON-*

2011 - 4th Biomedical Engineering International Conference, 2011, pp. 169–173, doi: 10.1109/BMEICON.2012.6172044.

[58] M. Goyal, N. D. Reeves, A. K. Davison, S. Rajbhandari, J. Spragg, and M. H. Yap, "DFUNet: Convolutional Neural Networks for Diabetic Foot Ulcer Classification," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 4, no. 5, pp. 728–739, Sep. 2018, doi: 10.1109/TETCI.2018.2866254.

[59] V. N. Shenoy, E. Foster, L. Aalami, B. Majeed, and O. Aalami, "Deepwound: Automated Postoperative Wound Assessment and Surgical Site Surveillance through Convolutional Neural Networks," in *Proceedings - 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018*, Jan. 2019, pp. 1017–1021, doi: 10.1109/BIBM.2018.8621130.

[60] L. Alzubaidi, M. A. Fadhel, S. R. Oleiwi, O. Al-Shamma, and J. Zhang, "DFU_QUTNet: diabetic foot ulcer classification using novel deep convolutional neural network," *Multimed. Tools Appl. 2019 7921*, vol. 79, no. 21, pp. 15655–15677, Jun. 2019, doi: 10.1007/S11042-019-07820-W.

[61] B. A. Pinero, C. Serrano, J. I. Acha, and L. M. Roa, "Segmentation and classification of burn images by color and texture information," *J. Biomed. Opt.*, vol. 10, no. 3, p. 034014, 2005, doi: 10.1117/1.1921227.

[62] M. Goyal, N. D. Reeves, S. Rajbhandari, and M. H. Yap, "Robust Methods for Real-Time Diabetic Foot Ulcer Detection and Localization on Mobile Devices," *IEEE J. Biomed. Heal. Informatics*, vol. 23, no. 4, pp. 1730–1741, Jul. 2018, doi: 10.1109/JBHI.2018.2868656.

[63] M. Goyal, N. D. Reeves, S. Rajbhandari, N. Ahmad, C. Wang, and M. H. Yap, "Recognition of ischaemia and infection in diabetic foot ulcers: Dataset and techniques," *Comput. Biol. Med.*, vol. 117, p. 103616, Feb. 2020, doi: 10.1016/J.COMPBIOMED.2020.103616.

[64] S. Thomas, "Medetec Wound Database: stock pictures of wounds." http://www.medetec.co.uk/files/medetec-image-databases.html (accessed Jun. 09, 2021).

[65] Darrenl and Tzutalin, "LabelImg," *Git code*, 2015. https://github.com/tzutalin/labelImg (accessed Mar. 23, 2020).

[66] "ImageNet Classification." https://pjreddie.com/darknet/imagenet/ (accessed May 21, 2020).

[67] "Image Classification on ImageNet." https://paperswithcode.com/sota/image-classification-on-imagenet (accessed May 22, 2020).

[68] J. Hui, "Object detection: speed and accuracy comparison (Faster R-CNN, R-FCN, SSD, FPN, RetinaNet and YOLOv3)." https://medium.com/@jonathan_hui/object-detection-speed-and-accuracy-comparison-faster-r-cnn-r-fcn-ssd-and-yolo-5425656ae359 (accessed May 22, 2020).

[69] D. P. Yadav, A. Sharma, M. Singh, and A. Goyal, "Feature extraction based machine learning for human burn diagnosis from burn images," *IEEE J. Transl. Eng. Heal. Med.*, vol. 7, pp. 1–7, 2019, doi: 10.1109/JTEHM.2019.2923628.

[70] A. Abubakar, H. Ugail, and A. M. Bukar, "Can Machine Learning Be Used to Discriminate Between Burns and Pressure Ulcer?," *Adv. Intell. Syst. Comput.*, vol. 1038,

pp. 870–880, Sep. 2019, doi: 10.1007/978-3-030-29513-4_64.

[71] C. Serrano, B. Acha, T. Gómez-Cía, J. I. Acha, and L. M. Roa, “A computer assisted diagnosis tool for the classification of burns by depth of injury,” *Burns*, vol. 31, no. 3, pp. 275–281, May 2005, doi: 10.1016/J.BURNS.2004.11.019.

[72] C. Serrano, R. Boloix-Tortosa, T. Gómez-Cía, and B. Acha, “Features identification for automatic burn classification,” *Burns*, vol. 41, no. 8, pp. 1883–1890, Dec. 2015, doi: 10.1016/J.BURNS.2015.05.011.

[73] B. Acha, C. Serrano, I. Fondon, and T. Gomez-Cia, “Burn depth analysis using multidimensional scaling applied to psychophysical experiment data,” *IEEE Trans. Med. Imaging*, vol. 32, no. 6, pp. 1111–1120, 2013, doi: 10.1109/TMI.2013.2254719.

[74] O. Despo *et al.*, “BURNED: Towards Efficient and Accurate Burn Prognosis Using Deep Learning,” 2017.

[75] “Burns BIP_US Dataset,” *Biomedical image processing (bip) group from the signal theory and communications department (university of seville, spain) and virgen del roc~Ao hospital (seville, spain)*. http://personal.us.es/rboloix/Burns_BIP_US_database.zip.

[76] I. Goodfellow *et al.*, “Generative Adversarial Networks,” *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Jun. 2014, doi: 10.48550/arxiv.1406.2661.

[77] F. H. Foomani *et al.*, “Synthesizing time-series wound prognosis factors from electronic medical records using generative adversarial networks,” *J. Biomed. Inform.*, vol. 125, p. 103972, Jan. 2022, doi: 10.1016/J.JBI.2021.103972.

[78] C. A. NILSSON and M. VELIC, “Classification of Ulcer Images Using Convolutional Neural Networks,” 2018.

[79] B. Rostami, D. M. Anisuzzaman, C. Wang, S. Gopalakrishnan, J. Niezgoda, and Z. Yu, “Multiclass wound image classification using an ensemble deep CNN-based classifier,” *Comput. Biol. Med.*, vol. 134, p. 104536, Jul. 2021, doi: 10.1016/J.COMPBIOMED.2021.104536.

[80] S. Sarp, M. Kuzlu, E. Wilson, U. Cali, and O. Guler, “A Highly Transparent and Explainable Artificial Intelligence Tool for Chronic Wound Classification: XAI-CWC,” no. January, pp. 1–13, 2021, doi: 10.20944/preprints202101.0346.v1.

[81] H. Wannous, Y. Lucas, and S. Treuillet, “Enhanced assessment of the wound-healing process by accurate multiview tissue classification,” *IEEE Trans. Med. Imaging*, vol. 30, no. 2, pp. 315–326, Feb. 2011, doi: 10.1109/TMI.2010.2077739.

[82] F. J. Veredas, R. M. Luque-Baena, F. J. Martín-Santos, J. C. Morilla-Herrera, and L. Morente, “Wound image evaluation with machine learning,” *Neurocomputing*, vol. 164, pp. 112–122, Sep. 2015, doi: 10.1016/J.NEUCOM.2014.12.091.

[83] F. Veredas, H. Mesa, and L. Morente, “Binary tissue classification on wound images with neural networks and bayesian classifiers,” *IEEE Trans. Med. Imaging*, vol. 29, no. 2, pp. 410–427, Feb. 2010, doi: 10.1109/TMI.2009.2033595.

[84] H. Wannous, S. Treuillet, and Y. Lucas, “Supervised tissue classification from color images for a complete wound assessment tool,” *Annu. Int. Conf. IEEE Eng. Med. Biol. - Proc.*, pp. 6031–6034, 2007, doi: 10.1109/IEMBS.2007.4353723.

[85] R. Mukherjee, D. D. Manohar, D. K. Das, A. Achar, A. Mitra, and C. Chakraborty, "Automated tissue classification framework for reproducible chronic wound assessment," *Biomed Res. Int.*, vol. 2014, 2014, doi: 10.1155/2014/851582.

[86] H. Wannous, Y. Lucas, and S. Treuillet, "Efficient SVM classifier based on color and texture region features for wound tissue images," *Med. Imaging 2008 Comput. Diagnosis*, vol. 6915, p. 69152T, Mar. 2008, doi: 10.1117/12.770339.

[87] H. Wannous, S. Treuillet, and Y. Lucas, "Robust tissue classification for reproducible wound assessment in telemedicine environments," *J. Electron. Imaging*, vol. 19, no. 2, p. 023002, Apr. 2010, doi: 10.1117/1.3378149.

[88] S. Zahia, D. Sierra-Sosa, B. Garcia-Zapirain, and A. Elmaghraby, "Tissue classification and segmentation of pressure injuries using convolutional neural networks," *Comput. Methods Programs Biomed.*, vol. 159, pp. 51–58, Jun. 2018, doi: 10.1016/J.CMPB.2018.02.018.

[89] V. Rajathi, R. R. Bhavani, and G. W. Jiji, "Varicose ulcer(C6) wound image tissue classification using multidimensional convolutional neural networks," *Imaging Sci. J.*, vol. 67, no. 7, pp. 374–384, Oct. 2019, doi: 10.1080/13682199.2019.1663083.

[90] H. Nejati *et al.*, "Fine-Grained Wound Tissue Analysis Using Deep Neural Network," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2018-April, pp. 1010–1014, Sep. 2018, doi: 10.1109/ICASSP.2018.8461927.

[91] E. Pasero and C. Castagneri, "Leg Ulcer Long Term Analysis," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10362 LNCS, pp. 35–44, 2017, doi: 10.1007/978-3-319-63312-1_4.

[92] G. Blanco *et al.*, “A superpixel-driven deep learning approach for the analysis of dermatological wounds,” *Comput. Methods Programs Biomed.*, vol. 183, p. 105079, Jan. 2020, doi: 10.1016/J.CMPB.2019.105079.

[93] R. Niri, H. Douzi, Y. Lucas, and S. Treuillet, “A Superpixel-Wise Fully Convolutional Neural Network Approach for Diabetic Foot Ulcer Tissue Classification,” in *In Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event*, 2021, pp. 308–320.

[94] F. Chollet, “Transfer learning & fine-tuning,” *Keras*. https://keras.io/guides/transfer_learning/ (accessed Jul. 02, 2021).

[95] “ImageNet.” https://image-net.org/update-mar-11-2021.php (accessed Apr. 04, 2022).

[96] B. Coetzee, R. Roomaney, N. Willis, and A. Kagee, “Body Mapping in Research,” *Handb. Res. Methods Heal. Soc. Sci.*, pp. 1237–1254, Jan. 2019, doi: 10.1007/978-981-10-5251-4_3.

[97] M. WILSON, “Understanding the basics of wound assessment,” *Wounds Essentials*, vol. 2, pp. 8–12, 2012.

[98] P. Krajcik, M. Antonic, M. Dunik, and M. Kiss, “PixelCut – PaintCOde.” https://www.paintcodeapp.com (accessed Jun. 15, 2021).

[99] J. Jonassaint and G. Nilsen, “The Application Factory – Body Map Picker.” https://github.com/TheApplicationFactory/BodyMapPicker (accessed Jun. 15, 2021).

[100] U. of Bristol, “Clickable bodymap,” *Bristol Medical School: Translational Health Sciences*. https://www.bristol.ac.uk/translational-health-

sciences/research/musculoskeletal/orthopaedic/research/star/clickable-bodymap (accessed Jun. 15, 2021).

[101] A. Slapšinskaitė, R. Hristovski, S. Razon, N. Balagué, and G. Tenenbaum, "Metastable Pain-Attention Dynamics during Incremental Exhaustive Exercise," *Front. Psychol.*, vol. 0, no. JAN, p. 2054, Jan. 2017, doi: 10.3389/FPSYG.2016.02054.

[102] M. Molenda, "Original Anatomy Mapper." https://anatomymapper.com (accessed Jun. 15, 2021).

[103] F. Chollet, "The Functional API," *Keras*. https://keras.io/guides/functional_api/ (accessed Jun. 18, 2021).