Impact of Quality of Image Database on AI Performance in Skin Cancer Detection

Yixuan Li

School of Computer Science / Department of Statistics

The University of Auckland

Supervisor: Prof. Patrice Delmas

A dissertation submitted in partial fulfillment of the requirements for the degree of Master of Professional Studies in Data Science / Master of Data Science, The University of Auckland, 2022.

Abstract

Skin cancer is one of the most lethal cancers worldwide, accounting for 1/3 of the total cancer diagnosed every year. Predominantly in Caucasian people, melanoma is the most dangerous form among all of skin cancers, characterized by the small proportion but the highest death rate, ~75% of all skin cancers death. Due to the continuous depletion of the ozone layer and 10-20% annual increase of skin cancer diagnosis, the automated, fast and effective early detection is in high demand to increase the five years' survival rate of patients and to find the effective therapeutical approaches on time. With the development of booming AI technology, convolutional neural networks have been widely studied and applied in skin cancer detection, with the peer level comparable to the dermatologists' diagnosis.

This work was done in Kahu AI to investigate the connection of AI database with the performance of AI CNN-based model. Several Python programs were developed to track, analyse and manage the modification of the AI database. Several major issues were investigated, including redundancies and miss-labelling of images in the database. After cleaning the databases, the performance of the patented Kahu AI model was compared, such as, accuracy, sensitivity, specificity of the detection of binary and multi-label classification. The results show at 95% sensitivity, the accuracy and specificity of Melanoma detection are increased ~ 1% after the cleaning. To systematically study this effect, I artificially introduced noise in the AI database with 5%, 10% and 20% duplicates of cross-labelling & repetitive-labelling, and applied the EfficientNet-B4 CNN model on ISIC Melanoma detection dataset with / without artificially introduced noise to compare the performance of AI. The results show that cross-labelling would severely reduce the training accuracy while the impact to validation accuracy is not obvious. On the other hand, AI performance remains resistant to the noise of repetitive-labelling for both training and validation, as well as testing datasets.

Table of Contents

Chapter 1
Introduction

Skin cancer [1] is among the most dangerous forms of cancer worldwide, accounting for 1/3 of the total cancer diagnosed every year. Among all skin cancer types, melanoma is one of the most aggressive and resistant to treatment types of human cancers. Although it takes only 5% of skin cancer [2], but accounts for 75% of all deaths from skin cancer, for example, more than 10,000 deaths annually in Australia and the United States. It has the highest mortality rate of all dermatological cancers, and is one of the most common cancers in young adults under 30, especially in young women. However, only 1 out of 20-40 lesions in the skin would be diagnosed as melanoma. The end stage of melanoma only has a 15% of survival rate. Hence, early diagnosis is crucial for cancer survival rates.

1.1 Cause of skin cancer and bias of population

Due to ozone depletion [3] caused by global warming and environmental contamination, the solar ultraviolet radiation reaching the surface of the earth is increasing over years, which could damage biological tissues by mutating DNA in cells. It is the major cause of an increase in the occurrence of skin cancer, especially in Australia and New Zealand. Skin cancer happens predominantly among Caucasian people. It represents around 35 to 45% of all skin diseases in Caucasians [4]. White Caucasian people naturally lack melanin pigment in their skin. Melanin [5] can give skin its colour and protect the skin from harmful UV light by acting as a physical barrier and as an absorbent filter that reduces the penetration of UV through the epidermis. Wide-range screening of melanoma [6] in this population is crucial to reduce death, however, the lack of funding is a major problem. The cost of treating skin cancer is tremendous, for example, the estimated annual cost of treating skin cancers in the U.S. is ~$8.1 billion, about $4.8 billion for non-melanoma skin cancers and $3.3 billion for melanoma [1]. Early detection of all skin cancer is essential to improve the morbidity and survival of patients.

1.2 Skin structure and major types of skin cancer

As shown in Figure 1 [7], normal skin has three layers:

> **-The epidermis**, the outermost layer of skin. It consists of keratinized, stratified squamous epithelium, which is generally made of four or five layers of epithelial cells, depending on its location in the body. In the thinner part of skin, like eyelid, there are 4 layers of epitheliums. Keratinocytes are predominant cells in the epidermis which manufacture and store the protein keratin. Keratin provides mechanical strength, resistance to chemicals in our skin, and acts as a waterproof barrier. A special type of cell called melanocyte which produces the pigment melanin is located in the deepest layer of the epidermis. Melanin gives the colour of our skin and hair tone. Besides melanocytes, there is a single layer of basal cells composed of columnar keratinocytes adjacent to the basement membrane, acting as the precursors of keratinocytes. All of the keratinocytes are produced from this single layer of cells. Squamous cells are multi-layered flat cells located near the surface of the skin that died and fall of continuously as new cells form.

-The dermis, the connective tissue layer of the skin that is about 0.5 to 5 mm thick, and located under the epidermis. It contains tough fibrous connective tissue, hair follicles, and sweat glands. It provides strength and elasticity to the skin due to the presence of collagen and elastin fibres. The dermis consists two layers, the papillary layer and the reticular layer. The capillary layer is composed of loose fibrous connective tissue with lymph and blood capillaries, sensory nerve fibres, etc... They supply nutrients to epidermis. The reticular layer of the dermis is deeper, thicker, and stronger, which is composed of three-dimensional network of thick bundles of collagen fibres interacting with the network of elastic fibres, forming a dense fibrous uniformed connective tissue.

-The deeper subcutaneous tissue (**hypodermis**), a continuation of the dermis, is made of fat and layers of loose fibrous connective tissue. This subcutaneous tissue plays an important role in the body, acting as a heat insulator and a storage place for nutrients, hormones, and vitamins.

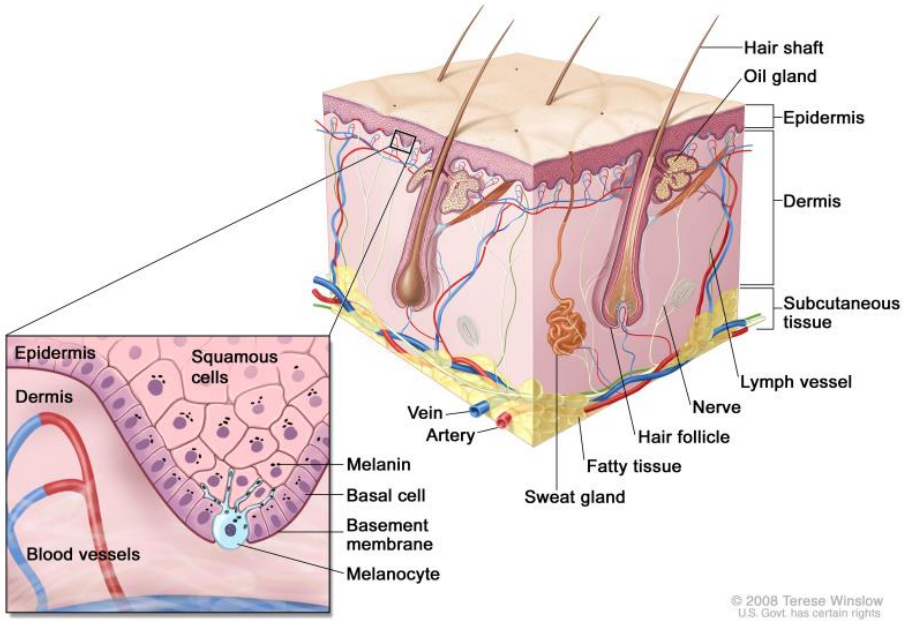


Figure 1. Schematic view of normal skin.

The three most common types of skin cancer are **basal cell carcinoma (BCC), squamous cell carcinoma (SCC), and melanoma** [1].

BCC is the most common but least serious type of skin cancer. It causes a lump, bump, or lesion to form on the epidermis layer. This is mostly due to over-exposure to the sunshine. Treatment to remove cancer from your skin leads to a positive outcome.

The second most common skin cancer is SCC. SSC is a type of skin cancer caused by an overproduction of squamous cells in your epidermis. Usually, SCC appear frequently on sun overexposed areas, such as the face, neck, and hands. They can also form scars or chronic skin sores. SCC cancer can only affect the top layer of the skin or can spread beyond the top layer of your skin. If not treated on time, it can be metastatic, which means the tumour would spread to other parts of your body beyond your skin.

Melanoma is not as common as BCC or SCC but is the most dangerous type of skin cancer. If it kept untreated or delayed till a late stage, melanomas are more likely to spread to organs except the skin through lymph nodes, blood or other tissues, making them difficult to treat and potentially causing death. Over-exposure to sunlight is the major cause of melanoma. Early melanomas are easily to be treated, with the 5 years survival rate of >90%, compared to metastasis to distant organs (20% survival rate) or lymph nodes (63% survival rate) at the late stage. Melanoma can look like moles, scaly patches, sores, or protruding bumps. One can use the American Academy of Dermatology's "ABCDE" memory device [8] to learn if a spot on your skin might be melanoma:

Asymmetry: One half does not match the other half. The shape is not uniform.
Border: The edges are not smooth.
Colour: The colour is mottled and uneven, with shades of brown, black, grey, red, or white.
Diameter: The spot is greater than the tip of a pencil eraser (6.0 mm).
Evolving: The spot is newly formed or continuously changes in size, shape, or colour.

Some melanoma may not show ABCDE signs. However, any unusual bumps, rashes or changes in your skin, surrounding area, or in any existing moles could be the source of melanoma, it should be checked by a dermatologist.

There are also some rare types of skin cancer, such as Merkel's carcinoma, Kaposi's sarcoma, and dermatofibrosarcoma protuberans (DFSP). Although Merkel's carcinoma is less than 1% of malignant, it is the second major cause of death from skin cancer second to melanoma. DFSP is less common than Merkel's carcinoma and least metastatic.

1.3 Detection of skin cancer

Skin cancer rates are doubling around the world every 10-20 years and dermatology posts are often unfilled due to significant dermatologist shortages. Moreover, non-specialists, like general practitioners, have only 20-70% diagnostic accuracy. Hence, automatic fast-reliable skin cancer detection method is in high demand. Many dermatological imaging researchers believe that diagnosis of skin melanoma can be automated based on certain physical features and colour information that are characteristic of the different categories of skin cancer [9], such as the vertical thickness, three-dimensional (3D) size and shape, and colour of the lesion, shape in the boundary of the lesion, and the appearance of pigmentation.

Traditionally, the initial diagnosis step of skin cancer is through visually detecting suspected spots, either through self-examinations or skin clinical examination, i.e., clinical screening. The morphology, such as colour, texture, size, and shape of the suspected area get examined with bare eyes. Clinically various imaging methods can be used to facilitate skin examination. As a typical example, a dermatoscopy [10] which magnifies the skin lesion (a mole or area of skin with an unusual appearance in comparison with the surrounding skin) using a bright light source, is commonly used by most skin cancer specialists. If the skin examined is suspicious of cancerous transformation, skin biopsy with histopathology is the gold standard for differential diagnosis of skin cancers. However, a biopsy which requires the removal of cells or skin samples from the lesion area or area around for histological examinations is an invasive and sometimes painful method. Skin biopsy is comparatively expensive, time-consuming, and typically followed by scar formation afterwards. Repetitive tissue sampling is impossible for patients with several

suspicious lesions. Moreover, repetitive clinical presentations are challenging to patients, because it can cause stress and anxiety to them. In addition, mistaking one skin cancer for another can lead to the wrong treatment being employed or lead to a delay in effective treatment planning, in some extreme cases, for example, a false negative diagnosis of melanoma could delay the surgical removal, risking cancer spreading to other organs in the body, and causing possibly death of patients. Besides the dermatoscopy, there are other non-invasive imaging procedures can be used as well, such as confocal laser scanning microscopy (CLSM)[11] , especially reflectance confocal microscopy(RCM)[12][13], magnetic resonance imaging (MRI) [14], optical coherence tomography (OCT) [15], high-frequency ultrasound(HFUS) [16], Terahertz pulsed imaging [17], and Raman Spectroscopy [18] [19]. I did a comparative table of all above mentioned methods with their merits / demerits as listed in Table 1. Due to the higher water content of cancerous tissues, the refractive index and absorption coefficient of normal tissues and cancerous tissues are different, which can be captured by some imaging techniques as shown in Table 1. Molecular imaging of skin cancer assisted with imaging contrast agents [20], CT-scan, and XR ray are usually for advanced-stage skin cancer which helps doctors to see if cancer has spread or come back, but not for early-stage diagnosis. Some genetic-information retrieval techniques are available to distinguish melanomas from naevi, for example, adhesive tape stripping [21] was developed using a 17-gene genomic biomarker, which shows 100% sensitivity and 88% specificity. Based on high-resolution dermatoscopic images or other imaging methods, there have been many techniques developed over the last decades for the early detection of skin cancer, including computer-assisted diagnosis (CAD) [22], machine learning algorithms, such as decision tree algorithms [23] Bayesian classifiers [24], support vector machines [25][26], and deep neural networks.

Imaging methods	Principle	Advantage	Disadvantage
Dermatoscopy[10]	a bright light source	non-invasive, conventional clinical diagnostic method. Dermatoscopy is more accurate than visual inspection alone, esp. when interpreted with the patient present.	Surface morphological examination only.
Reflectance confocal microscopy (RCM) [11][12][13]	using a point source of light, which is tightly focused on a specific point in the tissue. The light is reflected back by certain tissue structures due to variations of refractive indices within the skin. melanin, hydrated collagen, and keratin are highly reflective skin components.	Non-invasive, fast, real-time imaging at cellular level resolution (lateral resolution is 0.5–1 μm, axial resolution is ~ 3–5 μm); can perform optical sectioning. Average sensitivity of detecting melanoma is 92.7%, specificity 78.3%; for BCC, sensitivity 91.7%, specificity 91.3%. It is a very promising technique.	With a restricted depth of penetration 200-300 μm; lack of nuclear details compared to histopathology, small viewing field; medical staffs require extensive training.
Magnetic resonance imaging (MRI) [14]	MRI makes use of the magnetic properties of protons or other nuclei to generate high-resolution images.	Accurate cross-sectional information, exquisite soft tissue contrast with good spatial resolution	Expensive, low Sensitivity and slow acquisition of images. Sometimes requires contrast agents.

		(25-100 μm), without limit of penetration depth; providing both diagnostic information and the outcome of therapeutic intervention.	
Optical coherence tomography (OCT)[15]	A microscopic imaging technique, magnifying the surface of a skin lesion using near-infrared light.	Non-invasive, acquiring cross-sectional, high-resolution (2-20 μm) imaging of structures below the tissue surface (1-3 μm). OCT helps to reduce the false positive rate.	costly and requiring specialist training; not able to distinguish between BCC, cSCC, and melanoma.
High-frequency ultrasound (HFUS) [16]	measuring sound wave reflections from body tissues with transducer frequencies >=20 MHz.	Non-invasive, easy to use, availability of 3D imaging; penetration depth 6-7mm; with the potential to distinguish melanoma and basal cell carcinoma from other harmless lesions.	Poor resolution 50-300 μm
Terahertz pulsed imaging (TPI) [17]	Terahertz radiation is an electromagnetic wave with a frequency that lies in 0.1 to 10^{13}Hz range.	Both morphological and functional information of skin cancer, with resolution 20-200 μm, serving as a research and potentially therapeutic instrument.	Not fully validated, frequency-dependent; with penetration depth (commonly <1 μm), generating heat; can cause DNA damage at high frequency.
Raman Spectroscopy [18][19]	Using a diode laser light source; Raman signals correlate with the molecular vibrations of various tissue biomolecules.	Non-invasive, molecular level characterization, spatial resolution < 1 μm, binary classification sensitivity 95% - 99%, specificities 15% - 54%, ROC of melanoma 0.823-0.898	With penetration depth 300 μm, poor quantitative repeatability, expensive technique

Table1. Comparison of seven major imaging techniques for skin cancer detection.

Convolutional Neural Networks (CNNs) [27] are a class of AI deep learning techniques, most commonly used to analyse visual images with the capability to automatically detect significant features without human supervision. The major components of CNN include layers, activation functions, and hyperparameters. The structure of CNN layers, inspired by neurons in human and animal brains, contains one input layer, one output layer, and multiple hidden layers. The hidden layers include convolutional layers, pooling layers, and fully connected layers. Convolutional layers extract key information from images to a feature map; pooling layers reduce data dimension in order to increase the computational efficiency; fully connected layers connect every neuron in one layer to every neuron in another layer and output classifications. The convolution kernels, which are several random matrices containing certain target patterns

within the input image, is the key element of a convolutional layer. Various choices of kernels could achieve different image operations: feature identification, edge detection, blur, sharpening, etc. The activation function is a function of transformation that maps the input signals into output signals that are required for the neural network to function. Popular types of activation functions include linear activation, Sigmoid functions, Rectified linear units (ReLU), Exponential Linear Unit, and Softmax. Hyperparameters include filter kernel, batch size, padding, learning rate, optimizers, etc.

CNN-based skin cancer diagnostic tools involve customizing computer algorithms through a process called training to learn from data formed by predefined features. There are various well-known architectures of CNN models that can be used to train the computer, such as GoogleNet [28], ResNets [29] [30], LeNet [31], AlexNet [32], VGGNet [33], and EfficientNet [34]. These architectures are key structures in building deep learning algorithms. With the fast development of AI technologies, it is hardly surprising that they are being used to assist in diagnosing skin cancer and suggesting courses of action because AI-based methods are considered to be relatively cheap, easy to use, and accessible. Many studies have proved that AI is capable of classifying skin cancer with a level of competence comparable to dermatologists [2][35-39]. I had done a survey on recent researches using various CNN models compared with human raters. Table 2 shows a detailed comparison of these kinds of human vs. AI comparative studies. We can see AI accuracies are either equivalent to or outperforming human raters in most skin cancers, except the rare types of skin cancers. It indicates the era of AI-assisted clinical skin cancer diagnostics has come.

CNN architecture	**Dataset**	**Performance**
GoogleNet Inception v3 CNN [2]	129,450 clinical images for training for total 2,032 classes of skin diseases; 127,463 for training, validation, 1,942 for testing	72.1 ± 0.9% of CNN accuracy for 3-class model vs. 66% of human raters, 55.4 ± 1.7% of CNN accuracy in 9-class model vs. 53-55% of human raters. For binary classification of melanoma and keratinocytic carcinoma, CNN outperformed human raters in both sensitivity and specificity.
A combined CNN-based classification (cCNN) model, InceptionResNetV2, InceptionV3, Xception, ResNet50, training on both dermoscopic and close-up images, only best model chosen [35]	7,895 dermatoscopic images and 5,829 close-up images, 2072 test images, 51 classes	average accuracy of CNN 0.742 at 95% CI higher than or equivalent to human raters 0.695; specificity of CNN about the mean level of human raters (51.3%), sensitivity of the cCNN (80.5%) > human raters (77.6%), not good on rare skin lesions due to shortage of training.
ResNet50 CNN [36]	Total 2169 melanomas and 18,566 atypical nevi, 12,378 training dermatoscopic images, 2 classes	At sensitivity of 74.1% (human raters level), CNN has specificity of 86.5% . At a mean specificity of 60% (human raters' level), CNN has mean sensitivity of 87.5%. Both outperformed human raters.

Google's Inception v4 CNN [37]	>100.000 dermatoscopic pictures for training, 100 for testing, including 25% melanomas and 75% benign melanocytic nevi	Level-2 dermatologists achieved a mean sensitivity 88.9%, specificity 75.7%. At sensitivity 88.9%, CNN specificity is 82.5% outperforming human rating. AUC of CNN (0.86) also higher than human rating (0.79)
EfficientNet-B3 [38]	25,773 clinical images, 80/20 training/validation split, 10 classes, 3507 testing images	The overall accuracy of CNN is 78.45% vs. 0.73 for human rater for 10-class classification. Binary classification sensitivity 89.56% (assisted with CNN) vs. 83.21% (human), specificity 87.90% (assisted with CNN) vs. 80.92% (human)
Fully CNN similar to U-Net architecture combined with support vector machine [39]	ISIC-2016 dataset,720 training, 180 validation, and 379 testing images	76% (CNN) accuracy vs. 70.5% (human) and 62% (CNN) vs. 59%(human) specificity at equivalent sensitivity of 82% of human raters

Table 2. A list of recent AI studies in comparison with dermatologists' diagnosis.

In order to search for the best model to boost the accuracy of AI training and predictions, various CNN architectures had been employed to test the accuracy of skin cancer detection [40-55]. I collected recently published papers in this field. A detailed comparison of CNN model performance in these papers is shown in Table 3.

Architecture of CNN	**Dataset**	**Performance**
Resnet34 CNN with pre-processing of images [40]	ISIC 2019, 25,331 clinical-skin disease images for training & validation,8238 for testing	Sensitivity 0.8331, Accuracy 0.92
ResNet18, ResNet50, ResNet101 and ResNet152 [41]	In total 6,599 images from Kaggle, 1500 malignant and 1800 benign images for training, 1500 malignant and 1799 benign for testing, 2 classes	Accuracy 86.34% for ResNet18 model, 88.78% for ResNet50, 89.09% for ResNet101 and 89.65% for ResNet152
ResNet-101 and Inception-v3 [42]	2437 training, 660 test, 200 validation images, 2 classes	Accuracy 84.09% with ResNet-101, Accuracy 87.42% with Inception-v3
VGG-16, VGG-19, MobileNet, ResNet50 [43]	ISIC skin cancer datasets, 2,300 images, 9 classes	VGG-16 accuracy 0.4194, VGG-19 accuracy 0.4109, Mobile Net accuracy 0.4078, ResNet50 accuracy 0.3202
Resnet 50 with XGboost model structure optimization and leaky RELU activation function [44]	ISIC DATASET 2019	Accuracy 98.34% and precision 97.35%

ResNet50 CNN [45]	ISIC HAM10000 dataset 12378 images, two classes, melanoma and atypical nevi	The mean sensitivity and specificity 89.4% and 64.4%, At the same sensitivity, the CNN exhibits a mean specificity of 68.2%, equivalent to dermatologists' diagnosis.
Full-resolution convolutional networks (FrCN) used to segment the boundaries of skin lesions from dermatoscopic images before passing images to ResNet 50 CNN [46]	ISIC2018 dataset: training, validation, and testing with 2,000, 150, and 600 images,3 classes	overall accuracy of 94.03%, Jaccard similarity index of 77.11% via FrCN, overall accuracy of prediction 81.57%, F1-score 75.75%.
Designed CNN compared with Resnet50, InceptionV3, and Inception Resnet with ESRGAN image pre-processing step [47]	ISIC2018 dataset, 3533 skin images, 2 classes, malignant and benign	designed CNN accuracy 83.2%, Resnet50 (83.7%), InceptionV3 (85.8%), Inception Resnet (84%) models
VGG16 with preprocessing using U-net model, compared with Resnet, Xception, DenseNet [48]	ISIC DATASET 2594 images, 2075 training and 519 validation images	Accuracies are 0.75(Resnet), 0.76(Xception), 0.84 (DenseNet), 0.87(ensembled VGG16) and precisions are 0.71 (Resnet), 0.81 (Xception), 0.83 (DenseNet), 0.84 (ensembled VGG16)
VGG16, VGG19, MobileNet, and InceptionV3 with transfer learning [49]	HAM10000 dataset,10015 dermatoscopic images, train, validation, test ratio, 80:10:10, 7 classes	Accuracy achieved on VGG16, VGG19, MobileNet, and InceptionV3 is 87.42%, 85.02%, 88.22%, and 89.81%,
VGG 19 CNN and Transfer Learning [50]	HAM10000, 1920 training, 480 validations, 600 test images, 3 classes	training and testing accuracy were 0.985 and 0.975. training and testing loss were 0.099 and 0.119
EfficientNet-b4 modified with seven auxiliary classifiers to each of the intermediate layer [51]	13,603 images,14 classes	overall sensitivity of 93.38 ± 0.08%, specificity of 94.85 ± 0.05% in 14-class model, AUC 98.5% outperformed over original EfficientNet, ResNet-101, Inception-v3
pre-trained GoogLeNet Inception v3 [52]	7,192 dermoscopic images	overall classification accuracy of 81.49% in multiclass model, 77.02% in two-class classification
self -developed deep CNN model [53]	4,867 clinical images, 14 classes	overall accuracy 93.4% in differentiating benign and malignant conditions
self - developed deep learning system (DLS) including various number of deep CNN and shallow module for metadata [54]	14,021 training images, 3,756 for validation, 26 classes	0.71 and 0.93 top-1 and top-3 accuracies, top-3 accuracy 0.93 and average top-3 sensitivity 0.83 for 26 conditions
EfficientNet-B3 with Neural Architecture Search (NAS) and Model Quantization technique to reduce model	HAM10000, total 10015 images, 8000 training, 3 classes	Accuracy of Base Model 0.93114, Accuracy of Base Model with Quantization Aware Training 0.86527

size to 1/4 of original model [55]		

Table 3. A list of recent skin cancer research using different CNN architectures.

Different CNN architectures have their own special principles [56][57]. AlexNet is considered the first model that aroused interest in CNNs when it won the ImageNet challenge in 2012. It utilizes large and small-sized filters on the initial (5x5 and 11x11) and last layers (3x3) for feature extraction. ResNet employs identity-based skip connections or shortcuts to switch to certain layers of residual learning architecture. DenseNet employs depth-wise and cross-layer dimensions which ensures maximum data flow between the layers in the network. Xception uses depth-wise separated convolution layers. VGG is characterized by a pyramidal shape and composed of a series of convolutional layers followed by pooling layers with the pooling layers contributing to the narrower shape of the architecture. GoogleNet utilizes multiscale filters within the layers and employs a system of split, transform, and merge processes. Inception V3 & V4 employ deep feature hierarchies and multilevel feature representation. Inception-ResNet incorporated both Inception blocks and Resnet architectures. EfficientNet performs compound scaling, which means scaling in all three dimensions (width, depth, and resolution) while maintaining a balance between all dimensions of the network.

Despite the improvement of state-of-art CNN models and pre-processing techniques, as well as the activation functions, and tuning of hyperparameters, one fundamental question is how the quality of the image database would affect the AI performance. In the hyper-speedy life of general practitioners and clinical dermatologists, it is common to face tens of patients every day. Mistakes could happen every day when updating patients' information in the system. Besides, a patient can visit multiple clinics at the same time, and the information on enrolment could be completely varied from time to time. These could cause chaos in the image database and its connection with the associated patients and their diagnosis.

1.4 Purpose of this study

This work is conducted by Kahu.ai Ltd which is developing a software product to facilitate the general practitioners with the diagnosis of skin lesions suspicious of malignancy or melanoma. It also employs a specific dermatoscope to facilitate AI diagnostics. The software is an AI algorithm based on convolutional neural networks with a patented architecture that improves the diagnostic accuracy. The purpose of this project is to understand the relationship between the image dataset quality on the diagnostic performance of Kahu's AI. In order to achieve the goal, I assessed the quality of image database, which contains over 800,000 digital dermatoscopic images, including MAC (macro images), MIC-POL (micro-polarized images), MIC-NON (non-polarized images). The redundancies and mislabelling of images were detected and corrected. AI performance before and after the database cleaning was compared. Furthermore, I mimicked similar effects on the ISIC melanoma database with an EfficientNet convolutional neural network (CNN) in order to systematically analyse the impact of image databases on AI performance.

Chapter 2
Methodologies

2.1 Introduction of AI image database

The AI image database is a file folder in the Linux system which contains over 800K digital dermatoscopic images with their labels of classes/subclasses that are represented as the names of the folders in the system. The folder of each class label contains 3 subfolders, MAC, MIC-POL, or MIC-NON, representing 3 different types of images. There are 3 levels of classes (as shown in table 4): the first level (level L0 classification) contains only two classes, benign and malignant; the second level (Level L1 classification) contains four classes, benign and 3 major subclasses of malignant skin cancers, melanoma, intraepidermal carcinoma (IEC or iec) and non-melanoma skin cancer (NMSC or nmsc); the third level (level L2 classification) has in total 39 subclasses, which are the more detailed subclasses for benign and malignant skin lesions. As an example, nmsc in level L2 classification would be divided into 7 subtypes as shown in Table 4. In a csv file used in our analysis, the typical filename is displayed as the format.

benign_benign_keratinocytic_solar lentigo_/MIC.POL/MMAU24311.45030170.MIC.20181018144429615.POL.jpg

The format of the filenames is displayed as L0 level_L1 level_ L2 level_/type of image/ jpeg image name (including global ID, patient ID, image type, etc. information)

Level	Description	
L0	Benign	Malignant
L1	Benign	Melanoma , iec, nmsc
L2	Dermatofibroma, keratinocytic, nevus acral, nevus agminate, nevus atypical, nevus benign, nevus blue, nevus compound, nevus congenital, nevus dermal, nevus encockarde, nevus halo, nevus involuting regressing, nevus irritated, nevus junctional, nevus lentiginous, nevus papillomatous, nevus reed nevus, nevus spitzoid, nevus traumatized, nevus ungual, other, other ephelides, other ink spot lentigo, other lentigo, other melanosis, vascular	iec: actinic keratosis , iec: scc in situ, melanoma: lentigo melanoma, melanoma: melanoma, melanoma:nodular melanoma, nmsc:bcc basal cell carcinoma, nmsc:bcc nodular basal cell carcinoma, nmsc:bcc pigmented basal cell carcinoma, nmsc:bcc recurrent basal cell carcinoma, nmsc:bcc superficial basal cell carcinoma, nmsc:scc keratoacanthoma, nmsc:scc squamous cell carcinoma

Table 4. Description of 3 level classification of skin cancers.

2.2 Tracking, analysing, and managing the modification of the AI databases

A Python program was developed for tracking the changes in the image database. This image database served as the base of AI training, validation, and testing purpose. The modifications of the AI database could be the addition of new images, the deletion of old images, and the modifications of labelling (path) of images in the database, which included the renamed images or renaming a complete folder or subfolders (diagnosis).

The input files are two CSV files that indicate the current and previous status of the AI database. The pseudocode of the procedure can be found in Appendix. The result would be uploaded to a MySQL database and a log record would be updated in a logfile, which included the information of the user, name of the application, and summary of modification.

2.3 Tracking the quality of AI databases

When updating patients' information in the clinical system, many mistakes could happen. In order to track the individualism of images and their associated diagnostic information. We proposed SHA hashing techniques on the jpeg images. The SHA-256 algorithm, also named Secure Hash Algorithm 256, was created by the National Security Agency in 2001 as a successor to SHA-1. SHA-256 is a patented cryptographic hash function. It can generate a value of 256 bits long. The '256' in the name stands for the length of the final hashed digits. Being irrespective of the size of plaintext/cleartext/images to be hashed, the final hash value will always be 256 bits. This technique was normally used for cyber security, for example, digital signature verification, password hashing, SSL handshake, or integrity checks. It has been more popular for image encryption nowadays [58].

A Python program was developed for hashing the image files into 256 bits long hash codes and uploading/storing them in a SQL database in order to monitor the status of the AI database and keep track of all image-files in the system. The original purpose of hashing images was to:

1, track the current total jpeg files of the AI database.
2, track the unwanted man-made errors causing a modification of the AI database. If there are filenames modified accidentally, the image filenames and their hash codes are not correlated anymore.
3, archive the records of images in the AI database for long-term monitoring.

The program would scan over the designated folders and subfolders, encrypted all jpeg files in the folders, and output a CSV file including filename, file path, and hashcodes of 256 bits length. The resulting file would be further uploaded onto a MySQL database. In addition, a corresponding log record would be uploaded in logfile in SQL database as well, which included the information of the username, the specific python application, and the summary of modifications. Furthermore, through running SQL query, the resulting new table (imagehash_n) would be compared with the old table (imagehash_n-1) which contained the previous status of the AI database. N was the total number of imagehash tables, which were shown as imagehash1, imagehash2, imagehash3…in SQL database. Any result of the comparison showing records with the same hash codes but different filenames and / or file paths would be filtered and output in an error.csv file. The pseudocode of image-hashing software can be found in appendix.

2.4 Evaluation of AI performance

AI results were generated using Kahu's patented CNN-based algorithm before and after the modification of the AI database. A python program was developed to analyse the AI performance before and after the modification. After running the AI algorithm, the raw AI results were generated as a CSV file with all probabilistic scores for each class of skin lesions (as columns) and for every image file (as rows). It contained filenames/paths and a list of AI scores ranging from 0 to 1 (not including 0 and 1) in association with their pre-diagnosed class labels L0, L1, and L2 for all image files in the AI database. For each dataset, there were about 30-40K records.

The corresponding prevalence tables, Receiver Operating Characteristic (ROC) curves, and Confusion Matrices (CM) were generated. Furthermore, the results produced before and after the cleaning of AI databases were compared.

2.5 Testing AI performance using EfficientNet-B4 model

Conventional CNN model fitting uses more layers (depth) so that the richer and more complex features can be captured, or wider networks (width) for more fine-grained features, or higher resolution images that in theory can capture more details of the features. Instead of finding the best architecture, EfficientNet CNN [34] model starts with a relatively small baseline model and gradually scale it for the optimum settings of depth, width and resolution, instead of finding the best architecture. By using a compound coefficient Ø, one can uniformly scales network width, depth and resolution in this way [34]:

$$\text{Depth: } d = \alpha^{\emptyset} \text{ ; Width: } \omega = \beta^{\emptyset} \text{ ; Resolution: } r = \gamma^{\emptyset}$$
$$\text{such that: } \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2, \text{ given that } \alpha \geq 1, \beta \geq 1, \gamma \geq 1 \quad (1)$$

where α, β, γ are constants that can be determined by a small grid search to find the optimum scaling parameters. Intuitively, φ is a coefficient that controls how many more resources are available for model scaling, while α, β, γ specify how to assign these extra resources to network width, depth, and resolution respectively. So far, EfficientNet is one of the best models for CNN modelling with higher accuracy and 5-8 times faster than the conventional CNN models.

To mimic the AI database in Kahu, I used ISIC Melanoma Detection Dataset on Kaggle [59], which contains 2000 training images, 150 validation images, 600 test datasets. It has 3 classes of skin lesion photos, melanoma, nevus (a benign type of skin lesion, non-cancerous and most easily to be misdiagnosed with when melanoma presents), and seborrheic keratosis (another benign type of skin lesion, non-cancerous and sometimes mixed with melanoma in clinical diagnosis).

The original dataset has three directories: train, valid, and test. Each directory contains three subdirectories: melanoma, nevus, and seborrheic keratosis. Each of these subdirectories contains images for that specific class. The total images for melanoma are 374 in training dataset, 117 in testing dataset, 30 in validation dataset; for nevus, 1372 in training dataset, 393 in testing dataset, 78 in validation dataset; for seborrheic keratosis, 254 in training dataset, 90 in testing dataset, 42 in validation dataset. To replicate the different errors in the AI database, I randomly sampled and duplicated images according to the ratios of 5%, 10%, and 20% of the total number of training images, while maintained class ratios for 3 lesion types. The datasets remained the same for validation and testing datasets. For the repetitive labelling study, the images were copied and pasted into the same directory under the name copy (original name).jpg. For the cross-labelling study, the redundant images of class 'melanoma' were copied and pasted into the folder of 'nevus', while the duplicated images of the class 'nevus' and 'sborrheic_keratosis' were moved into the folder of 'melanoma'. Therefore, the same images were both labelled as nevus and melanoma, or seborrheic keratosis and melanoma.

Due to the efficiency (fast and less parameters comparatively) and high accuracy of EfficientNet CNN architecture as reported (table 3), in this study I applied Google's EfficientNet-b4 with pre-trained weights on ImageNet as the backbone of our CNN architecture and trained all datasets for 50 epochs using Adam optimizer and Relu activation function. Initially a global learning rate 0.001 was used and gradually decayed to 0.0001 at epoch 50. In the training process, each image

was resized to 380∗380 pixels in RGB channels, which was the optimized input size of EfficientNet-b4. For each epoch, each image would be rotated from −30◦ to 30◦ randomly, together with 50% probabilities for vertical and horizontal flipping. Image normalization was also applied to convert an input image into the optimum pixel values. The model was updated according to the three classifiers for Melanoma classification datasets. The accuracy and the loss of train, validation and test datasets over 50 epochs before and after the modification of the melanoma database were generated. The result was plotted for visual comparison.

Chapter 3
Results and Discussion

3.1 Tracking, analyzing, and managing the modification of the AI databases

Table 8 shows a sample of input CSV files (same format for old and new CSV files which represent the status of the current AI database and the earlier AI database) and the output result.csv of comparison after measuring the difference of two CSV files. The column of status showed what kind of modification was done to the particular image files: 'new' means the newly added image files; 'delete' means the deleted old photos; or label-changed images. There were 7 different situations in the label modification: L0 changed but L1 & L2 unchanged (shown as L0change); L1 changed but L0 and L2 unchanged (shown as L1change); L2 changed but L0 and L1 unchanged (shown as L2change); L0 and L1 changed but L2 unchanged (shown as L0L1change); L1&L2 changed but L0 unchanged (shown as L1L2change); L0 and L2 changed but L1 unchanged (shown as L0L2change); and L0, L1, and L2 all changed (shown as L0L1L2change) as displayed in the Table 5 lower table. The date shows the updated date of the result in SQL database.

name	label	is_train	data_type
benign_benign_keratinocytic_actinic cheilitis_/MIC.POL/VIC201506.15010232.MIC.20150122114450301.POL.jpg	benign:benign:keratinocytic	1	MIC.POL
malignant_melanoma_melanoma_/MIC.POL/MMAU31954.25170806.MIC.20191008100139584.POL.jpg	malignant:melanoma:melanoma	1	MIC.POL
benign_benign_nevus benign_/MIC.POL/MMNZ05192.14890477.MIC.20150225121517642.POL.jpg	benign:benign:nevus benign	1	MIC.POL
benign_benign_keratinocytic_solar lentigo_/MIC.POL/VIC107208.15260406.MIC.20150408152148975.POL.jpg	benign:benign:keratinocytic	1	MIC.POL
malignant_iec_actinic keratosis_common/MIC.POL/MMNZ05544.15010151.MIC.20170628120650948.POL.jpg	malignant:iec:actinic keratosis	1	MIC.POL
benign_benign_keratinocytic_seborrhoeic keratosis_common/MIC.POL/VSK101150.45210129.MIC.20150311124	unknown:benign:keratinocytic	1	MIC.POL
benign_benign_nevus compound_/MIC.POL/BTNY16964.24850598.MIC.20160209104816415.POL.jpg	benign:unknown:nevus compound	1	MIC.POL
benign_benign_nevus atypical/MIC.POL/MMAU01691.24670418.MIC.20181128164637528.POL.jpg	benign:benign:unknown	1	MIC.POL
benign_benign_nevus halo_/MIC.POL/WVL201968.24860506.MIC.20140415121517367.POL.jpg	unknown:unknown:nevus halo	0	MIC.POL
benign_benign_keratinocytic_seborrhoeic keratosis_common/MIC.POL/MMNZ51326.25330405.MIC.2018121816	unknown:benign:unknown	1	MIC.POL
malignant_iec_actinic keratosis_common/MIC.POL/DUNE01523.34900150.MIC.20141126122352944.POL.jpg	unknown:unknown:unknown	1	MIC.POL
malignant_nmsc_scc squamous cell carcinoma_/MIC.POL/MMNZ49186.31941613.MIC.20181004135153026.POL	malignant:nmsc:scc squamous cell carcinom	1	MIC.POL
malignant_melanoma_melanoma_/MIC.POL/MMNZ09844.15160516.MIC.20150608140135888.POL.jpg	malignant:unknown:unknown	1	MIC.POL
benign_benign_other_chrondrodermatitis_/MIC.POL/MMNZ02651.45350133.MIC.20141202083942981.POL.jpg	benign:benign:other	1	MIC.POL
benign_benign_keratinocytic_solar lentigo_/MIC.POL/NELS13059.35330125.MIC.20160516141300487.POL.jpg	benign:benign:keratinocytic	1	MIC.POL
malignant_nmsc_bcc basal cell carcinoma_/MIC.POL/QGRE00963.25920414.MIC.20150629114440771.POL.jpg	malignant:nmsc:bcc basal cell carcinoma	1	MIC.POL
benign_benign_nevus dermal_/MIC.POL/VRIN01383.15040290.MIC.20150805100311523.POL.jpg	benign:benign:nevus dermal	1	MIC.POL
malignant_iec_actinic keratosis_common/MIC.POL/MMPT02336.35050158.MIC.20200107145603280.POL.jpg	malignant:iec:actinic keratosis	1	MIC.POL
benign_benign_keratinocytic_solar lentigo_/MIC.POL/MMNZ01901.31490566.MIC.20141104112919950.POL.jpg	benign:benign:keratinocytic	1	MIC.POL

photo	L0	L1	L2	status	date	is_train	data_type
MMPT02336.35050158.MIC.20200107145603280.POL.jpg	malignant	iec	actinic keratosis	delete	8/07/2022	1	MIC.POL
MMNZ01901.31490566.MIC.20141104112919950.POL.jpg	benign	benign	keratinocytic	delete	8/07/2022	1	MIC.POL
QGRE00670.34821061.MIC.20161213143220160.POL.jpg	benign	benign	nevus atypical	delete	8/07/2022	1	MIC.POL
HAMI23223.62930505.MIC.20190415083051277.POL.jpg	benign	benign	nevus compound	delete	8/07/2022	1	MIC.POL
ASPF00293.30750770.MIC.20141210131339619.POL.jpg	malignant	nmsc	scc squamous cell car	delete	8/07/2022	1	MIC.POL
ASP701916.31470648.MIC.20150420105554212.POL.jpg	malignant	nmsc	bcc superficial basal c	new	8/07/2022	1	MIC.POL
MMAU03431.48670664.MIC.20191113094842750.POL.jpg	malignant	iec	actinic keratosis	new	8/07/2022	1	MIC.POL
MMNZ50140.15800370.MIC.20181127114816665.POL.jpg	malignant	melanoma	melanoma	new	8/07/2022	1	MIC.POL
TAUR18841.45180988.MIC.20131111154853777.POL.jpg	benign	benign	nevus atypical	new	8/07/2022	0	MIC.POL
VRIN00881.13000855.MIC.20160725115038786.POL.jpg	malignant	iec	actinic keratosis	new	8/07/2022	1	MIC.POL
MMNZ41032.13440396.MIC.20200115113800044.POL.jpg	unknown	iec	actinic keratosis	L0change	8/07/2022	1	MIC.POL
ROTO01368.15450150.MIC.20190910163552407.POL.jpg	unknown	unknown	actinic keratosis	L0L1change	8/07/2022	1	MIC.POL
MMNZ05192.14890477.MIC.20150225121517642.POL.jpg	unknown	unknown	unknown	L0L1L2change	8/07/2022	1	MIC.POL
ASP402813.26321799.MIC.20140521085756764.POL.jpg	malignant	unknown	bcc basal cell carcinor	L1change	8/07/2022	1	MIC.POL
MMAU19506.47570488.MIC.20180125101334593.POL.jpg	benign	unknown	unknown	L1L2change	8/07/2022	0	MIC.POL
MMNZ27361.25200584.MIC.20161025163618182.POL.jpg	malignant	unknown	unknown	L1L2change	8/07/2022	1	MIC.POL
MMAU31954.25170806.MIC.20191008100139584.POL.jpg	unknown	melanoma	unknown	L0L2change	8/07/2022	1	MIC.POL
VSK101150.45210129.MIC.20150311124120983.POL.jpg	benign	benign	unknown	L2change	8/07/2022	1	MIC.POL

Table 5. an example of data structure of old.csv/new.csv (up) and result.csv (down).

3.2 Investigation of AI database via image hashing

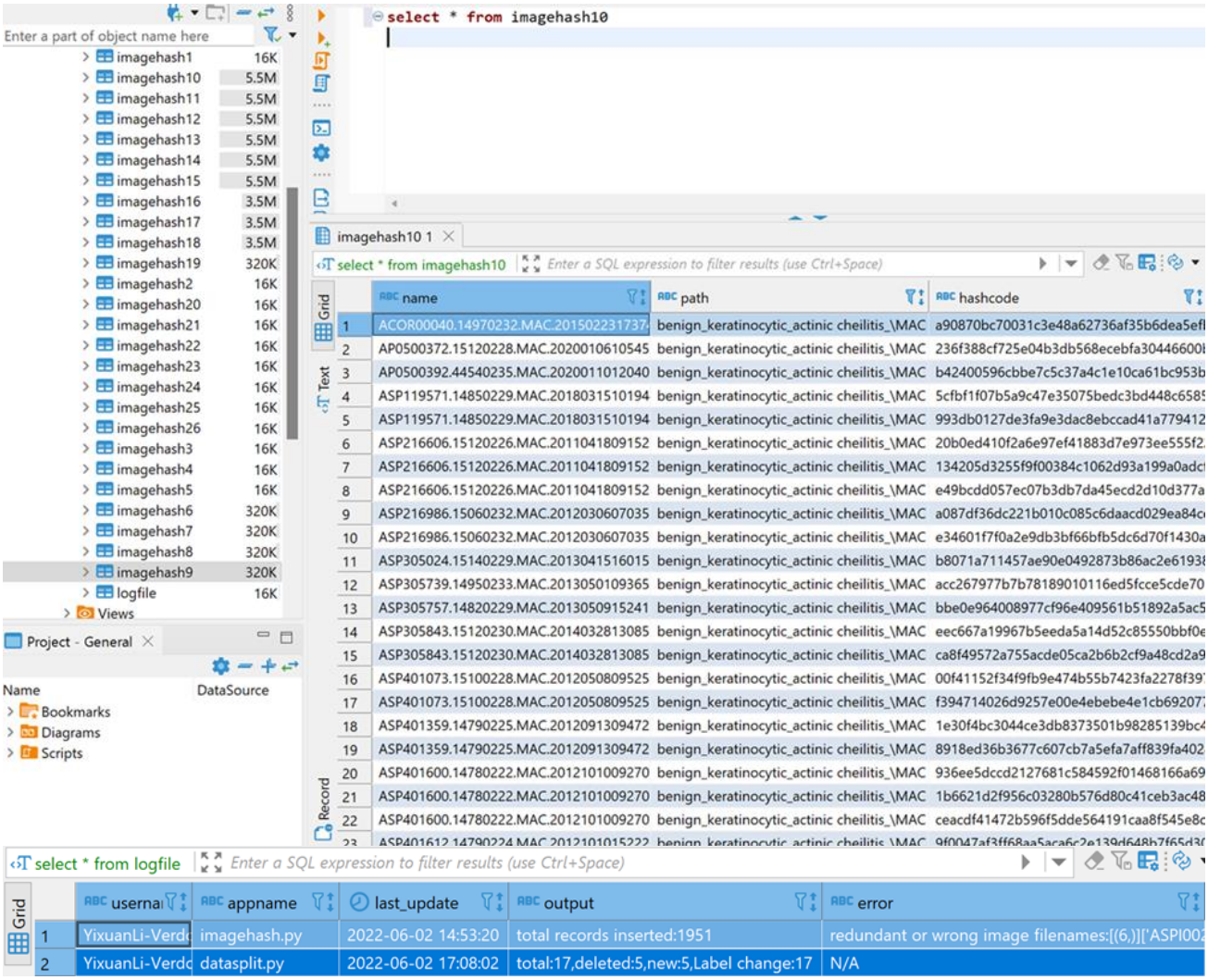


Figure 2. A screenshot of SQL database after uploading image-hashing result (up) and the records in the logfile (down).

Here (in Figure 2) shows a screenshot of the database status after importing the hashed image dataset and the updated logfile.

The total records in AI database were 835,214, including MAC, MIC.NON and MIC.POL images. A systematic analysis had been done based on image-hashing results which would reflect the current status of the AI database, as shown in table 6. I had investigated:

1, if redundant images existed with the same name and in the same location (label), named as “pure duplicates”.

2, if redundant images presented with the same name but under different directories, named as “duplicated image names” in the table.

3, if redundant images existed with different names but in fact identical jpeg files, named as “duplicated hash codes” in the table. I had double-checked these original photos that they are identical. The record showed the total number of the unique photos are 643,766 in the AI database.

4, If there were redundant images under both malignant and benign labelled directories. It was named as “cross-labelling” hereafter.

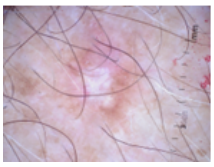
5, If there were redundant images under malignant or benign labelled directories but under different subtypes of malignancy or benign directories. It was named as “repetitive-labelling” hereafter. The result of image-hashing was analysed according to labels at levels of L0 / L1 classification in order to get detailed statistics of the redundant data distribution in system.

Description	Total image with labels	% in Total
Total	835,214	100.00%
Unique file name + labels	835,214	100.00%
No pure duplicates exist		
Duplicated image names	190,218	22.77%
Duplicated hash codes	191,448	22.92%
Difference: same images named with different filenames	1,230	0.15%
Unique images names	644,996	77.23%
Unique hashcode (absolute unique images)	643,766	77.08%
L0 labels analysis in duplicated filenames		
L0 - cross labelling of malignant and benign	1820	0.22%
L0 - repeative labelling of malignant or benign	188398	22.56%
subtotal:	190218	22.77%
L1 labels analysis in duplicated filenames		
L1 - cross labelling of benign:benign and malignant:iec	256	0.03%
L1 - cross labelling of benign:benign and malignant: melanom	1137	0.14%
L1 - cross labelling of benign:benign and malignant:nmsc	427	0.05%
L1 - cross labelling of malignant:iec and malignant:melanoma	16	0.00%
L1 - cross labelling of malignant:iec and malignant:nmsc	448	0.05%
L1 - cross labelling of malignant:melanoma and malignant:nm	66	0.01%
L1 - repeative labelling of benign:benign	187483	22.45%
L1 - repeative labelling of malignant:iec	177	0.02%
L1 - repeative labelling of malignant:melanoma	54	0.01%
L1 - repeative labelling of malignant:nmsc	154	0.02%
subtotal:	190218	22.77%

Table 6. Summary of AI database status and the detailed analysis of L0 and L1 labels in total 190,218 duplicated filenames.

The above table indicated there were no pure duplicates in the system. This was plausible. However, several major issues could be found:

1, The difference between the total number of duplicated image filenames and the total number of duplicated hashcodes was 1230 records in total. This indicated the identical jpeg images were named under different filenames as shown in Figure 3. These kinds of errors could be due to the mistakes on the clinical side and should be corrected.



ASP305456.31550590.MIC.20140428080042946.POL.jpg
ASP701991.31550590.MIC.20140428080042946.POL.jpg
ASPH03700.31550590.MIC.20140428080042946.POL.jpg

Figure 3. An example of an identical image named with three different filenames under path malignant_nmsc_bcc basal cell carcinoma_/MIC.POL.

2, More than 22% of duplicated filenames were found, among which 190,218 records were corresponding to 94,787 unique jpeg files, which gave the repetitive frequency ~ 2 per image file. It could be cause by the same jpeg files located at different file folders (labels). In order to understand the compositions of these labelling, I did an aggregation on L0 and L1 classification. Here showed the results.

1) L0 classification: benign and malignant

Two kinds of issues in L0 level classification were found, one was due to the same image files were both diagnosed as malignant and benign, i.e., cross-labelling. An example was shown in Table 7. The total cross-labelling occupied ~ 0.22% of the total data.

name	path	hashcode
@POD00004.14970729.MIC.20140503120707252.POL.jpg	malignant_melanoma_melanoma_/MIC	30fa4765b3cd19236d59cfef45faf25d01ecd4d15e95c3f41887e997d10bb846
@POD00004.14970729.MIC.20140503120707252.POL.jpg	benign_benign_vascular_telangiectasia_/benign_benign_nevus benign_/MIC.POL	30fa4765b3cd19236d59cfef45faf25d01ecd4d15e95c3f41887e997d10bb846
@POD00004.14970729.MIC.20140503120707252.POL.jpg	benign_benign_nevus atypical/MIC.POL	30fa4765b3cd19236d59cfef45faf25d01ecd4d15e95c3f41887e997d10bb846
@POD00004.14970729.MIC.20140503120707252.POL.jpg	benign_benign_nevus benign_/MIC.POL	30fa4765b3cd19236d59cfef45faf25d01ecd4d15e95c3f41887e997d10bb846
@POD00004.14970729.MIC.20140503120707252.POL.jpg	benign_benign_nevus congenital_/MIC.	30fa4765b3cd19236d59cfef45faf25d01ecd4d15e95c3f41887e997d10bb846

Table 7. L0 cross-labelling of both malignant and benign for the same image files. An example of file @POD00004.14970729.MIC.20140503120707252.POL.jpg was diagnosed with 4 different benign subclasses and 1 malignant melanoma type.

Another issue was the same jpeg files were diagnosed as different subclasses in the L0 classification, i.e., repetitive-labelling of either benign or malignant. Here showed a summary of both cases in L0 classification.

L0 distribution	Unique images	L0/benign	L0/malignant	Total
cross-labelling	852	966	854	1820
repetitive-labelling	93935	187483	915	188398
sum:	94787	188449	1769	190218

Table 8. Summary of L0 level cross-labelling and repetitive-labelling.

In summary, the Clear evidence of 0.22% cross-labelling and 22.6% of repetitive-labelling existed in the L0 level of classification.

2) L1 level classification analysis – four classes

To understand the majority of cross-labelling and what labels likely took the dominance in cross-labelling, a further analysis in L1 level was done. L1 level classification was composed of four subclasses, 'benign:benign', 'malignant:iec', 'malignant:melanoma', 'malignant:nmsc'. There were no files presented in more than two L1-class labels. Cross-labelling at L1 level had 6 cases: benign:benign cross-diagnosed with malignant:iec (denoted as benign X iec), benign:benign cross-diagnosed with malignant:melanoma (benign X melanoma), benign:benign cross-diagnosed with malignant:nmsc (benign X nmsc), malignant:iec cross-diagnosed with malignant:melanoma (iec X melanoma), malignant:iec cross-diagnosed with malignant:nmsc (iec X nmsc) and malignant:melanoma cross-diagnosed with malignant:nmsc (melanoma X nmsc).

L1 label distribution	unique images	benign:benign	malignant:iec	malignant:melanoma	malignant:nmsc	total
benign X iec	125	131	125			256
benign X melanoma	532	603		534		1137
benign X nmsc	195	232			195	427
iec X melanoma	8		8	8		16
iec X nmsc	224		224		224	448
melanoma X nmsc	33			33	33	66
benign only	93479	187483				187483
iec only	87		177			177
melanoma only	27			54		54
nmsc only	77				154	154
Sum:	94787	188449	534	629	606	190218

Table 9. L1 cross-labelling and repetitive-labelling in absolute numbers.

The result showed the majority of redundancy happened due to repetitive labelling of benign skin lesions of all subclasses (could be labelled as different benign subclasses). The second serious redundancy was the cross-labelling of benign:benign with malignant:melanoma.

3.3 AI results analysis

Through Kahu patented-CNN model prediction, the probabilistic (numeric) outputs were generated for each lesion as rows and all L0, L1, L2 classification types as columns. These numeric outputs represented:

(1) the relative likelihood that lesion characteristics may be associated with malignancy;
(2) the relative likelihood that the lesion was classified as one of the following categories: L0 classification – benign vs. malignancy; L1 classification - melanoma, nmsc, iec, and benign; L2 classification – benign containing 27 subtypes and melanoma containing 3 subtypes, nmsc containing 7 subtypes), iec containing 2 subtypes, therefore, in total 39 subtypes.

The plot of true positive rate (TPR), which is alternatively named sensitivity, recall or probability of detection, versus false positive rate (FPR), also named 1-Specifity, is called receiver operating characteristic (ROC) curve [60]. This method was originally developed for operators of military radar receivers dated back in 1941, thus named after it. It is widely used nowadays for binary classification in medicine, radiology, biometrics, forecasting of natural hazards, meteorology, model performance assessment, and other areas for tens of years. Its application is increasingly acknowledged nowadays especially in machine learning and data mining research. The area under the curve (AUC) [60], as an effective measure of accuracy, has been widely accepted as a meaningful interpretation for disease classification from healthy subjects. This curve plays a

central role in evaluating diagnostic ability of tests to distinguish the true status of the investigated subjects, finding the optimal cut off values (threshold), and comparing binary alternative diagnostic tasks when each task is performed on the same subject of classification. A confusion matrix, commonly used in machine learning, statistical classification [61], is a specific table layout that allows visualization of the performance of an algorithm in a supervised learning process. In our case as shown in Figure 4, each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class. The name originated from the fact that it makes it easy to see whether the outcome of the system is mislabelling two classes in comparison with the actual classes.

Table 10 listed hereunder is the prevalence table of the sample distribution of two AI results being compared. The composition of benign and malignant classes and subclasses are equivalent in both before and after cleaning of the database. Thus, the comparison is valid.

# of images	D5.2 before clearning	%	D5.3 after clearning	%
Number of samples	39366	100.0	31492	100
Total negatives	21650	55.0	17108	54.3
Total positives	17716	45.0	14384	45.7
IEC	9382	23.8	7320	23.2
NMSC	5779	14.7	4647	14.8
Total Melanoma	2555	6.5	2417	7.7

Table 10. Prevalence table of benign and malignant lesions of AI database before (version D5.2) and after cleaning (version D5.3) in absolute number and percentage of the total records.

3.3.1 Binary classification

To achieve the desirable sensitivity of detection, the sensitivity was set at 95% in order to get optimum cut-off value (threshold) for L0 classification, L1 melanoma vs. non-melanoma classification, and L2 melanoma vs. non-melanoma classification. Because of the high death toll of melanoma patients, it is preferably that the detection to be more sensitive with the trade-off of a high percentage of false positive. In Figure 4, ROC curves were compared by plotting Sensitivity against 1-Specificity for L0 label benign vs. malignant, L1-melanoma vs. non-melanoma (L1-mela ROC curve), as well as L2 melanoma vs. non-melanoma classification (L2-mela ROC curve). We obtained the confusion matrix of L0-benign vs. malignant, L1-melanoma vs. non-melanoma as well as L2-melanoma vs. non-melanoma classification before and after cleaning of AI database in absolute values. The major changes which had been done to the AI database before and after cleaning included,

1) removing 1,230 duplicated image filenames, i.e., the same jpeg photos but under different filenames as shown in Table 6;
2), some labelling modification of images in the system, for example, images with the same diagnostic types and diagnostic actions would be mapped into the same class label.

AUC is an indication of accuracy of the AI detection. It can be observed for L0 detection, AUC reduced slightly from 0.951 to 0.949; for L1 melanoma detection, accuracy increased from 91.8% (before cleaning) and 92.6% (after cleaning), and for L2 melanoma detection, accuracy increased from 91.8% to 92.5%.

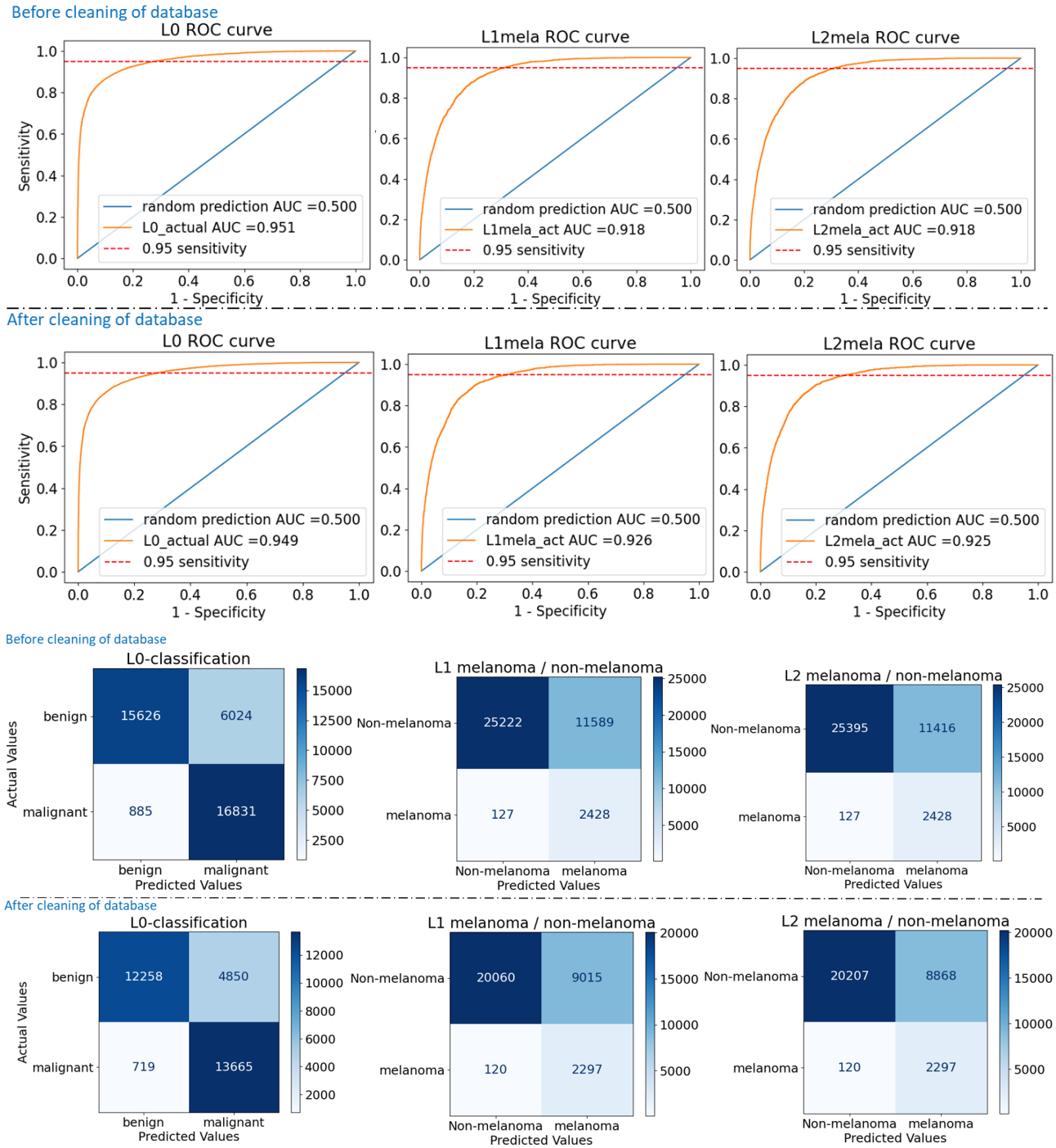


Figure 4. ROC curves and confusion matrices of L0, L1 melanoma (L1mela), L2 melanoma (L2mela) classification before and after database cleaning.

In the confusion matrix in Figure 4, the negative classes are set to be benign in L0 classification, Non-melanoma in L1 classification, and Non-melanoma in L2 classification. The positive classes are malignant in L1 classification, melanoma in L1 classification and melanoma in L2 classification. As shown in CM in Figure 4, the diagonal values indicate the true negative, abbreviated as TN (e.g., 15,626 in L0 classification before cleaning) and the true positive, abbreviated as TP (e.g., 16,831 in L0 classification before cleaning). The off-diagonal values are false positive values, abbreviated as FP (e.g., 6,024 in L0 classification before cleaning), and false negative values, abbreviated as FN (e.g., 885 in L0 classification before cleaning). Based on the

actual values in CM, the performance of AI algorithm was further evaluated from TP (true positive, if the outcome positive class from a prediction is the same as the actual positive class), FP (false positive, if the actual value is negative for predicted positive class), TN (true negative, when both the prediction outcome and the actual value are negative), FN (false negative, when the prediction outcome is negative while the actual value is positive), and their derivations [62]. Hereby the list of derivations is calculated based on TP, FP, TN, FN:

1, Sensitivity: TPR = TP / (TP+FN) (2)
2, Specificity: TNR = TN / (TN+FP) (3)
3, Precision: PPV = TP / (TP+FP) (4)
4, Distribution of lesion: 1/PPV (5)
5, Negative predictive value: NPV = TN / (TN+FN) (6)
6, False positive rate (FPR): FPR = 1-TNR (7)
7, Prevalence threshold: PT= $\frac{\sqrt{FPR}}{\sqrt{TPR}+\sqrt{FPR}}$ (8)
8, F1-score: $F1 = 2 \times \frac{PPV \times TPR}{PPV+TPR} = \frac{2TP}{2TP+FP+FN}$ (9)
9, Accuracy: $ACC = \frac{TP+TN}{TP+TN+FP+FN}$ (10)

The comparison of AI performance before and after database cleaning is listed in table 11. It can be observed that precision, accuracy improved slightly especially for melanoma classification. The distribution of lesion (equation 4) is an indication of how many lesion sites to be examined is likely to occur one true positive sample. It is slightly reduced after cleaning, an alternative indication of precision clinically. Negative predictive values (NPV, equation 5) indicate the ratio of true negative values to all sum of true negative and false negative (predicted) values. The prevalence threshold (equation 8) can help measure the validity of a screening test in real time, hereby not relevant to a diagnostic test, only for a reference. Accuracy (equation 10) measures the accuracy of the prediction of algorithm, which equals to the ratio of the total sum of true positive and true negative to the total sum of samples.

To increase the sensitivity of diagnosis, I set sensitivity at 95% and obtain the threshold for binary classification. The result would include large portion of false positive but it would also reduce the chance of patient's mis-diagnosis in melanoma as a preferred outcome of AI assisted diagnosis. About 70% specificity and 95% sensitivity is relatively quite favourable results in comparison with human raters in many literature studies (Table 2). F1-score (equation 9) measures the harmonic mean of sensitivity and precision, after cleaning, it has increased 4-5%. The accuracy of classification of both L1-melanoma and L2-melanoma increased 1%.

Key values	L0	L1 melanoma	L2 melanoma
Threshold TPR 95%	0.1543	0.0169	0.0172
Specificity	0.7218	0.6852	0.6899
AUC	0.951	0.918	0.918
After cleaning			
Threshold TPR 95%	0.1757	0.0225	0.0245
Specificity	0.7165	0.6899	0.695
AUC	0.949	0.926	0.925

Derivations	L0	L1 melanoma	L2 melanoma
Sensitivity(TPR)	0.95	0.95	0.95
Specificity(TNR)	0.72	0.69	0.69
Precision (PPV)	0.74	0.17	0.18
Distribuion of lesion(1/PPV)	1.36	5.77	5.70
Negative predictive value NPV	0.95	1.00	1.00
pevalence threshold(PT)	0.35	0.37	0.36
F1-score	0.83	0.29	0.30
Accuracy(ACC)	0.82	0.70	0.71
After cleaning			
Sensitivity(TPR)	0.95	0.95	0.95
Specificity(TNR)	0.72	0.69	0.70
Precision (PPV)	0.74	0.20	0.21
Distribuion of lesion(1/PPV)	1.36	4.93	4.86
Negative predictive value NPV	0.95	0.99	0.99
pevalence threshold(PT)	0.35	0.36	0.36
F1-score	0.83	0.34	0.34
Accuracy(ACC)	0.82	0.71	0.72

Table 11. Performance comparison of AI database before and after cleaning for binary classifications.

3.3.2 Four classes in L1 classification

L1 classification includes four subclasses, benign, malignant-IEC, malignant-melanoma and malignant -NMSC. The absolute and normalized values of each class can be found in Figure 5. Except melanoma predicted class inherited from L1-melanoma vs. non-melanoma binary classification, the rest of the predicted class of each image represent the highest AI scores for four classifiers values in L1 classes, denoted as following mathematically,

$$\text{classification} = \text{argmax}\left(\sum_{i=1}^{total\ number\ of\ Classifiers} AI\_Output(i)\right), \quad (11)$$

where AI_output(i) is a classification vector with dimension depending on the level of classification; for L1 level, the dimension is (1,4); for L2, it is (1,9); and the sum of the total classification vectors is an element wise summation. Argmax takes the maximum values out of the probabilistic scores of all classifiers and output the class label.

The actual classes are obtained from the class labels (name of the directory in AI database) as shown in filename.

In L1 Confusion Matrix, a row represents an instance of the actual class, whereas a column represents an instance of the predicted class. Consequently, the values of the diagonal elements represent the values of correctly predicted classes. The false classified off-diagonal elements are named the confusion, because they are mistakenly classified with another class by AI. When normalized over rows, the diagonal values show sensitivity (TPR) of each class or specificity (TNR) for other classes, for example, melanoma classification has 95% sensitivity and the corresponding specificity to benign, iec and nmsc are 0.45, 0.83, 0.71 before cleaning, versus 0.47, 0.83, 0.67 after cleaning. On the other hand, normalized over columns indicating precision (PPV) for each class with consistent actual and predicted class or negative predictive value (NPV) for other classes.

The results showed except the sensitivity for malignant-NMSC reduced from 0.71 to 0.67, the other values either remained the same or slightly improved, especially for the specificity of melanoma detection.

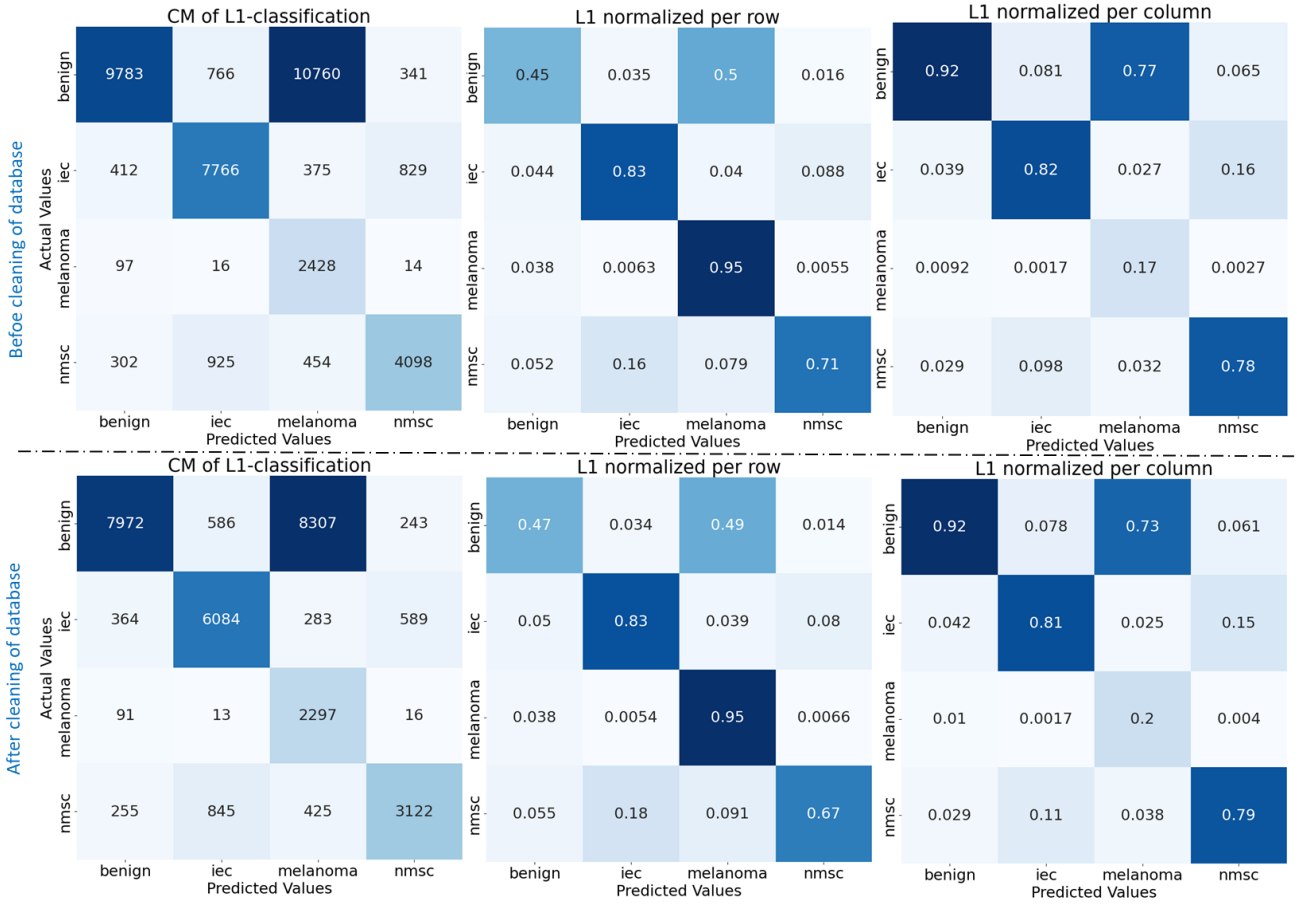


Figure 5. Confusion matrices of L1 labels classified into benign, IEC, melanoma and NMSC, top layer shows performance before database cleaning, bottom layer after database cleaning. Left images show the absolute number, middle images are normalized over rows, right images are normalized over columns

3.3.3 Nine classes in L2 classification

L2 labelling contains nine subclasses. Instead of classifying 39 classes, I grouped them into 9 major subtypes, including five subclasses in benign lesions: benign_vascular, benign_keratinocytic, benign_nervus, benign_other_dermatofibroma; 2 subclasses in malignant-

IEC; actinic_keratosis; scc_in_situ; malignant-melanoma; and 2 subclasses in malignant -NMSC: bcc and scc. The predicted classes were obtained same as in L1-classification, except the melanoma classification inherited from L1-melanoma binary classification, the rest of classifiers obtained according to the equation 11. The absolute and normalized values of each L2 class can be found in Figure 6. Except for the melanoma predicted class inherited from L2-melanoma vs. non-melanoma classification, the rest of the predicted class of each image represented the highest AI scores for nine groups of L2 classes.

Due to high sensitivity settings for melanoma, there was a large proportion of false positive rate of melanoma coming from benign_nervus. The specificity of L2 melanoma increased from 0.18 to 0.21. However, the sensitivity of bcc reduced from 0.73 to 0.68, since labelling change might have some comprehensive impact on the results, which was still under investigation.

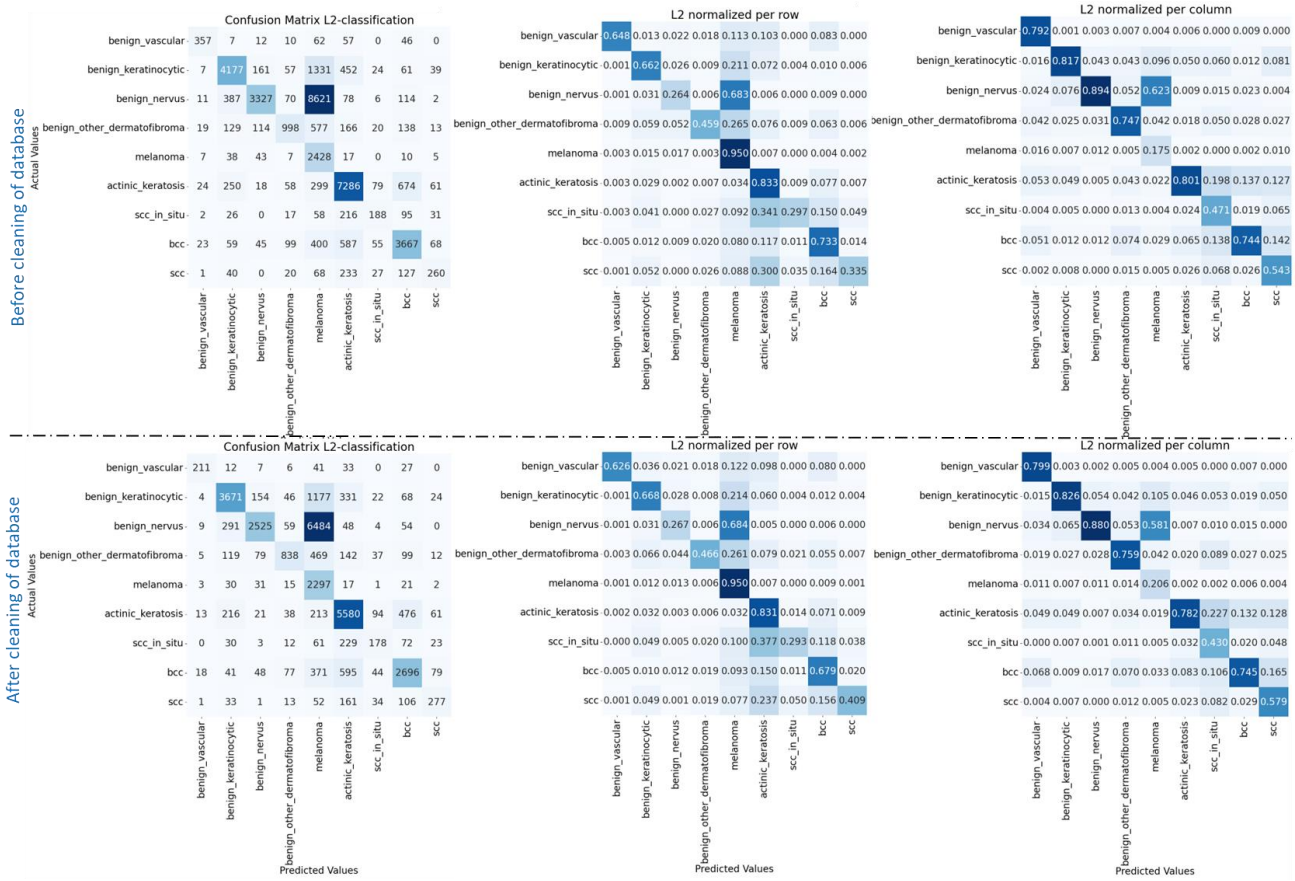


Figure 6. Confusion matrices of L2 classifications, showing performance before database cleaning (top), and after database cleaning(bottom). Left images showed the absolute number, middle images were ratios normalized over rows, right images were ratios normalized over columns.

3.4 Simulation of AI database with EfficientNet CNN model

To understand the compound relationship of database with the efficiency of CNN model prediction, the simulated AI databases were created for 5%, 10% and 20% redundancies of both cross-labelling and repetitive-labelling. The training and validation accuracy and loss were shown in Figure 7. Due to the limited time for this research and small size of training datasets, the model was probably slightly overfitted, which can be seen from the training loss was reduced over epochs but validation loss did not change after 5 epochs and remain fluctuating around 0.8.

However, from training accuracy (Figure 8) it can be observed a clear trend of diminishing accuracies with the increasing noise introduced in database, as well as increasing of the training loss. For repetitive labelling, AI modelling shows high fault tolerance, 5%,10%, even 20% repetitive-labelling do not show significant difference with the accuracy and loss of the original datasets. It shows a high fault tolerance of AI prediction in repetitive-labelling.

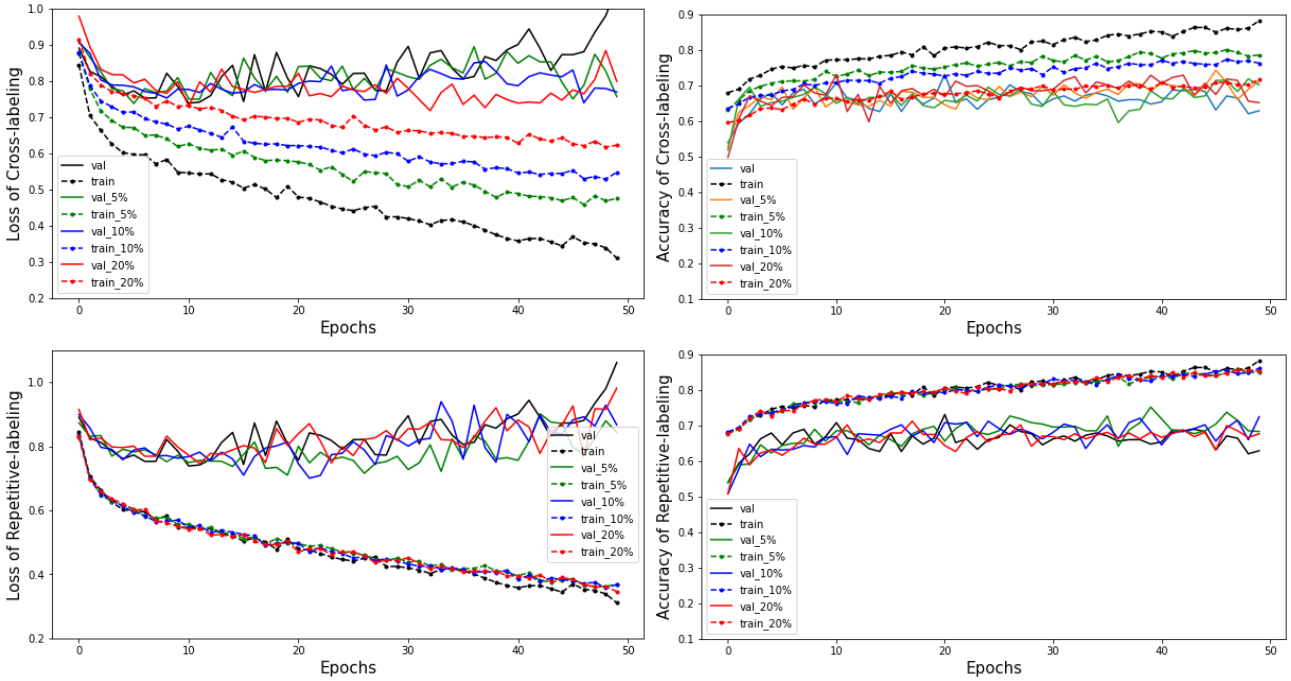


Figure 7. Accuracy and loss of Training and validation datasets. Upper two images showed the cross-labelling loss (left) and accuracy (right) of training (dashed line with markers) and validation (connected lines). Bottom two images showed repetitive-labelling loss (left) and accuracy (right) of both training (dashed line with markers) and validation (connected lines).

In cross-labelling, test accuracies were 0.673 (original dataset), 0.669 (5% noise), 0.633 (10% noise), 0.623(20% noise), and test losses were 0.783(original dataset), 0.801(5% noise), 0.839 (10% noise), and 0.845(20% noise). From results of AI performance on test datasets in Figure 8, for cross-labelling, it showed the tendency of declining test accuracy and inclining test loss with the increasing noise-ratio of cross-labelling in melanoma datasets. On the other hand, for repetitive-labelling AI performance was pretty resistant and tolerant to the database modification. In repetitive-labelling, test accuracies were 0.677 (original dataset), 0.664 (5% noise), 0.682 (10% noise), 0.637(20% noise), and test losses were 0.763(original dataset), 0.785(5% noise), 0.755 (10% noise), and 0.801(20% noise). It was likely to be no impact of repetitive-labelling to the performance of AI, or the impact is so minor that it could be balanced by the boosting effect of increasing size of the database, same as the results obtained from training and validation datasets in Figure 7.

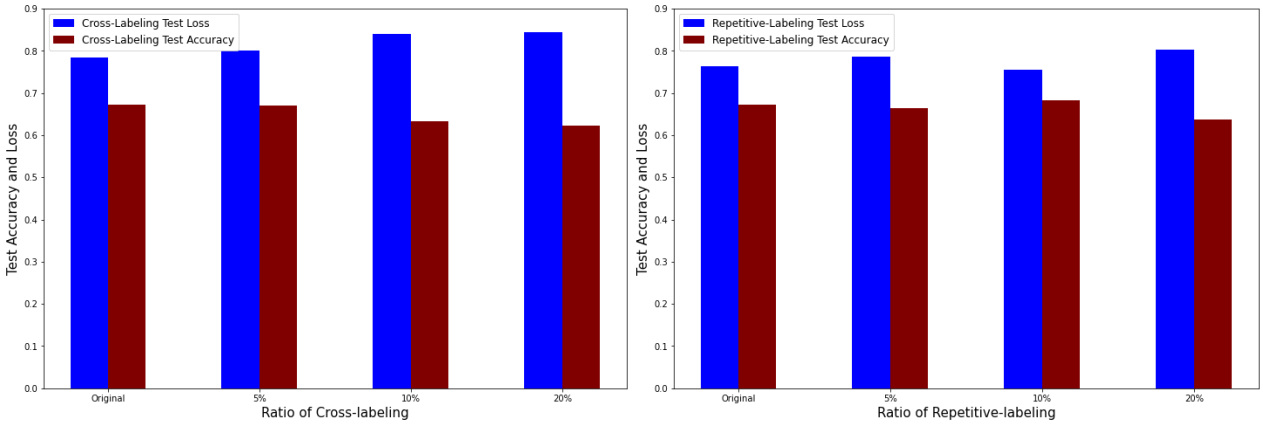


Figure 8. Accuracy and loss of Testing dataset. Left image showed the cross-labelling test loss (blue) and test accuracy (maroon). Right image showed the repetitive-labelling test loss (blue) and test accuracy (maroon).

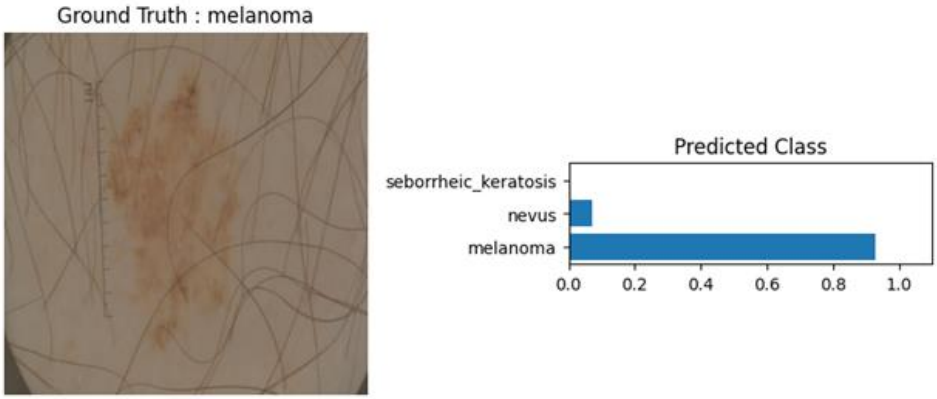


Figure 9. An example of melanoma image with the predicted values for different skin lesion classes.

In Figure 9, a typical image of melanoma skin lesion was presented with the predicted AI scores of 3 different classes that plotted in the horizontal bar plot. In this example, the predicted AI score for class melanoma is over 90%, and the predicted class is in consistency with the actual class.

Certainly, these were quite primary results. Due to limited amount of time, we did not do multiple testings to check the statistics range of accuracies and losses. These could only give a hint for the future research in this direction. It could be several ways to boost accuracy of AI detection, including increasing the size of the training database and balancing datasets of different skin lesion classes, as well as pre-processing images to remove hair or sharpen the images. As a result of this study, since repetitive-labelling did not impact the results, various augmented images could be used to boost datasets of rare skin lesions in skin cancer detection.

Chapter 4
Conclusions and Future Prospects

Skin cancer detection using AI technology is a very fast-growing field. Many products are released in the market, for example, Kahu SMARTI skin cancer detection system. I did a survey on the most common imaging techniques in skin cancer detection and reported the recent development on the CNN based skin cancer detection. It is encouraging that various studies based on AI deep learning had been proved to be at peer level or even superior to the dermatologists' diagnosis, although AI performance on the rare types of skin cancer is still inferior to human raters due to the lack of enough training datasets.

In this study, we did fundamental research on the influence of quality of database to AI performance. Image hashing is proved to be an effective way to track the uniqueness of the images despite its different image names. We investigated on Kahu AI database and found several issues including redundancy and mislabelling of jpeg files in the AI database. The mislabelling included both cross-labelling (different class labels) and repetitive-labelling (same class labels). After database cleaning, we compared AI results of binary, four-classes, and nine classes classification. The accuracy and specificity of melanoma detection all increased slightly, about 1%. To mimic the impact of AI database to the AI performance in the company, I did a systematic study on ISIC Melanoma Detection Dataset. I simulated 5%, 10% and 20% cross-labelling and repetitive-labelling redundancies (noise) in database. An EfficientNet-B4 CNN model was employed to facilitate this study. The results show cross-labelling would probably damage the AI prediction efficiency because training accuracy reduced with the increasing ratio of cross-labelling. Accuracy of test dataset showed minor reduction although the change of validation accuracy was not clearly visible. On the other hand, AI modelling showed high fault tolerance in repetitive labelling, although the reason was still unclear. It could be a mixture of mislabelling effect and model boosting effect on the increase size of database. So far, this research was probably the first investigation on the quality of database with the connection of AI performance.

Here is a list of possible future works in CNN based skin cancer diagnosis. The functionality of CNN algorithm is probably colour-biased, for example, for Asian people, AI should be trained with different skin lesion databases with dominant darker colour skin images. The class imbalance would potentially affect the result of AI as well. To boost accuracy of AI prediction, pre-processing techniques, like ESRGAN to enhance image resolution, or post-processing techniques, like XGboost to be a recognizer on the top level of the CNN network to produce AI outputs, would be adopted. Skin lesion segmentation methods via FCN, U-Net, and SegNet, FrCN, FrCN in cooperated with different CNN architecture outperformed all original methods according to literatures, which can be used to integrate with CNN architecture as a computer-aided diagnosis (CAD) system to assist clinical dermatology. Besides the technical improvement, an interesting trend of research direction is to incorporate the metadata of patients in the model to boost the performance of AI diagnosis.

References

[1] Skin Cancer Facts & Statistics. https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/ (accessed Nov. 14, 2022)

[2] Esteva, A., et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, 2017, https://doi.org/10.1038/nature21056

[3] Umar SA, Tasduq SA, "Ozone Layer Depletion and Emerging Public Health Concerns-An Update on Epidemiological Perspective of the Ambivalent Effects of Ultraviolet Radiation Exposure," *Front. Oncol.,* 12:866733, 2022, doi: 10.3389/fonc.2022.866733.

[4] Porcia T. Bradford, "Skin Cancer in Skin of Colour," *Dermatol Nurs.,* Vol. 21(4), pp.170–178, Jul-Aug. 2009, [online] Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2757062/

[5] Slominski RM, Sarna T, Płonka PM, Raman C, Brozyna AA, Slominski, "Melanoma, Melanin, and Melanogenesis: The Yin and Yang Relationship," *Front. Oncol.* 12:842496., Mar. 2022, doi: 10.3389/fonc.2022.842496.

[6] Johansson M, Brodersen J, Gøtzsche PC, Jørgensen KJ, "Screening for reducing morbidity and mortality in malignant melanoma," *Cochrane Database of Systematic Reviews*, Issue 6. Art. No.: CD012352, June 2019, doi: 10.1002/14651858.CD012352.pub2.

[7] "PDQ Adult Treatment Editorial Board, Melanoma Treatment (PDQ®): Health Professional Version, " PDQ Cancer Information Summaries.
https://www.ncbi.nlm.nih.gov/books/NBK66034.1/ (accessed Nov. 14, 2022)

[8] WHAT TO LOOK FOR: ABCDES OF MELANOMA. https://www.aad.org/public/diseases/skin-cancer/find/at-risk/abcdes (accessed Nov. 14, 2022)

[9] Azadeh Noori Hoshyar, A. Al-Jumaily and R. Sulaiman, "Review on automatic early skin cancer detection," *2011 International Conference on Computer Science and Service System (CSSS)*, pp. 4036-4039, 2011, doi: 10.1109/CSSS.2011.5974581.

[10] Dinnes J., et al., "Dermoscopy, with and without visual inspection, for diagnosing melanoma in adults." *Cochrane Database of Systematic Reviews*, Issue 12. Art. No.: CD011902, 2018. doi: 10.1002/14651858.CD011902.pub2.

[11] Ilie MA, et al., "Current and future applications of confocal laser scanning microscopy imaging in skin oncology," *Oncol Lett.*, Vol. 17(5), pp. 4102-4111, May 2019, doi: 10.3892/ol.2019.10066. Epub 2019 Feb 25.

[12] Ahlgrimm-Siess V, et al., "Confocal Microscopy in Skin Cancer," *Curr Dermatol Rep*., vol.7(2), pp. 105-118, 2018, doi: 10.1007/s13671-018-0218-9. Epub 2018 Apr 25.

[13] Levine A, Markowitz O, "Introduction to reflectance confocal microscopy and its use in clinical practice," *JAAD Case Rep.,* vol.4(10), pp.1014-1023, Nov. 2018, doi: 10.1016/j.jdcr.2018.09.019.

[14] Kawaguchi M, et al., "Magnetic Resonance Imaging Findings Differentiating Cutaneous Basal Cell Carcinoma from Squamous Cell Carcinoma in the Head and Neck Region," *Korean J Radiol*. vol.21(3), pp. 325-331, Mar. 2020, doi: 10.3348/kjr.2019.0508.

[15] Ferrante di Ruano L, et al., "Optical coherence tomography for diagnosing skin cancer in adults," *Cochrane Database of Systematic Reviews*, Issue 12. Art. No.: CD013189, Dec. 2018, doi: 10.1002/14651858.CD013189.

[16] Dinnes J, et al., " High-frequency ultrasound for diagnosing skin cancer in adults." *Cochrane Database of Systematic Reviews*, Issue 12. Art. No.: CD013188, Dec. 2018, doi: 10.1002/14651858.CD013188.

[17] Nikitkina AI, et al., "Terahertz radiation and the skin: a review," *J Biomed Opt*, vol. 26(4), pp. 043005, Feb. 2021, doi: 10.1117/1.JBO.26.4.043005.

[18] P. Santos, et al., "Improving clinical diagnosis of early-stage cutaneous melanoma based on Raman spectroscopy," *Br J Cancer*, Vol. 119, pp. 1339–1346, 2018. [online] Available: https://doi.org/10.1038/s41416-018-0257-9
[19] Harvey Lui; Jianhua Zhao; David McLean, Haishan Zeng," Real-time Raman Spectroscopy for In Vivo Skin Cancer Diagnosis," *Cancer Res* ,72 (10), pp. 2491–2500, May 2012, doi: 10.1158/0008-5472.CAN-11-4061.
[20] Hao Hong, Jiangtao Sun, Weibo Cai, "Anatomical and molecular imaging of skin cancer," *Clinical, Cosmetic and Investigational Dermatology*, vol.1, pp. 1–17, Oct. 2008, doi: 10.2147/ccid.s4249.
[21] Wachsman W, et al., "Noninvasive genomic detection of melanoma," *Br J Dermatol*, 164(4),pp.797-806, Apr. 2011, doi: 10.1111/j.1365-2133.2011.10239.x. Epub 2011 Mar 25.
[22] Ferrante di Ruano L, et al., "Computer-assisted diagnosis techniques (dermoscopy and spectroscopy-based) for diagnosing skin cancer in adults," *Cochrane Database of Systematic Reviews*, Issue 12. Art. No.: CD013186, Dec. 2018, doi: 10.1002/14651858.CD013186.
[23] M. Zorman, M. M. Štiglic, P. Kokol, and I. Malčić, "The limitations of decision trees and automatic learning in real world medical decision making," *J. Med. Syst.*, vol. 21, no. 6, pp. 403–415, Dec. 1997, doi: 10.1023/a:1022876330390.
[24] D. Ruiz, V. Berenguer, A. Soriano, and B. Sánchez, "A decision support system for the diagnosis of melanoma: A comparative approach," *Expert Syst. Appl*., vol. 38, no. 12, pp. 15217–15223, Nov. 2011. [online] Available: https://doi.org/10.1016/j.eswa.2011.05.079
[25] S. Gilmore, R. Hofmann-Wellenhof, and H. P. Soyer, "A support vector machine for decision support in melanoma recognition," *Exp. Dermatol*., vol. 19, no.9, pp. 830–835, Sep. 2010, doi: 10.1111/j.1600-0625.2010.01112.x. Epub 2010 Jul 11.
[26] Szyc Ł, Hillen U, Scharlach C, Kauer F, Garbe C, "Diagnostic Performance of a Support Vector Machine for Dermatofluoroscopic Melanoma Recognition: The Results of the Retrospective Clinical Study on 214 Pigmented Skin Lesions," *Diagnostics,* vol.9(3), pp.103, Aug. 2019, doi:10.3390/diagnostics9030103
[27] Alzubaidi, L, et al., "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *J Big Data*, 8, 53, Mar. 2021. [online] Available: https://doi.org/10.1186/s40537-021-00444-8 .
[28] Szegedy C, et al., "Going deeper with convolutions," *Proceedings of the IEEE conference on computer vision and pattern recognition*,2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.
[29] Wu S, Zhong S, Liu Y, "Deep residual learning for image steganalysis," *Multimed Tools Appl.,* 77(9), pp. 10437–10453, May 2018. [online] Available: https://doi.org/10.1007/s11042-017-4440-4 .
[30] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.
[31] LeCun Y, et al., "Learning algorithms for classifcation: a comparison on handwritten digit recognition," *Neural Networks Stat. Mech. Perspect,* pp. 261-276, 1995. [online] Available: https://www.researchgate.net/publication/2599424_Learning_Algorithms_For_Classification_A_Comparison_On_Handwritten_Digit_Recognition.
[32] Krizhevsky A, SI, Hinton GE, "Imagenet classifcation with deep convolutional neural networks," *Adv Neural Inf Process Syst*, pp. 1097–1105, 2012. [online] Available: https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html .
[33] Simonyan, K., and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *3rd International Conference on Learning Representations (ICLR 2015), Computational and Biological Learning Society*, 2015, pp. 1–14.

[34] Mingxing Tan, Quoc V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” *Proceedings of the 36th International Conference on Machine Learning*, 2019, PMLR 97.
[35] Tschandl P, et al., “Expert-Level Diagnosis of Nonpigmented Skin Cancer by Combined Convolutional Neural Networks,” *JAMA Dermatol.,* 1;155(1), pp. 58-65, 2019, doi: 10.1001/jamadermatol.2018.4378.
[36] T. J. Brinker et al., “Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task,” *European Journal of Cancer*, vol. 113, 2019, doi: 10.1016/j.ejca.2019.04.001.
[37] H. A. Haenssle et al., “Man against Machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists,” *Annals of Oncology,* vol. 29, no. 8, 2018, doi: 10.1093/annonc/mdy166.
[38] Wei Ba,Huan Wu,Wei W. Chen, et al., "Convolutional neural network assistance significantly improves dermatologists’ diagnosis of cutaneous tumours using clinical images," *European Journal of Cancer*, Vol. 169, pp. 156-165, 2022, [online] Available: https://doi.org/10.1016/j.ejca.2022.04.015 .
[39] N. C. F. Codella et al., "Deep learning ensembles for melanoma recognition in dermoscopy images," *IBM Journal of Research and Development*, vol. 61, no. 4, 2017, doi: 10.1147/JRD.2017.2708299
[40] N. Gouda and J. Amudha, "Skin Cancer Classification using ResNet," *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, 2020, pp. 536-541, doi: 10.1109/ICCCA49541.2020.9250855.
[41] M. Hossain, K. Sadik, M. M. Rahman, F. Ahmed, M. N. Hossain Bhuiyan and M. M. Khan, "Convolutional Neural Network Based Skin Cancer Detection (Malignant vs Benign)," *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)* , 2021, pp. 0141-0147.
[42] A. Demir, F. Yilmaz and O. Kose, "Early detection of skin cancer using deep learning architectures: resnet-101 and inception-v3," *2019 Medical Technologies Congress (TIPTEKNO),* 2019, pp. 1-4, doi: 10.1109/TIPTEKNO47231.2019.8972045.
[43] J. Liu, "VGG, MobileNet and AlexNet on Recognizing Skin Cancer Symptoms," *2022 3rd International Conference on Electronic Communication and Artificial Intelligence (IWECAI),* 2022, pp. 525-528, doi: 10.1109/IWECAI55315.2022.00107.
[44] R. Kaur and N. Kaur, "Improved Skin Cancer Detection Classification Residual Network Feature Engineering," *2021 International Conference on Computational Performance Evaluation (ComPE),* 2021, pp. 671-675, doi: 10.1109/ComPE53109.2021.9751930.
[45] Brinker TJ, et al., “A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task,” *Eur J Cancer.*, 111:148–54, Apr. 2019, doi: 10.1016/j.ejca.2019.02.005.
[46] M. A. Al-masni, M. A. Al-antari, H. M. Park, N. H. Park and T. -S. Kim, "A Deep Learning Model Integrating FrCN and Residual Convolutional Networks for Skin Lesion Segmentation and Classification," *2019 IEEE Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS),* 2019, pp. 95-98, doi: 10.1109/ECBIOS.2019.8807441.
[47] Gouda, W.; Sama, N.U.; Al-Waakid, G.; Humayun, M.; Jhanjhi, N.Z., “Detection of Skin Cancer Based on Skin Lesion Images Using Deep Learning,” *Healthcare (Basel)*, Vol. 10(7), pp.1183, June 2022, doi: 10.3390/healthcare10071183.
[48] M. Kumar and S. M. P. Gangadharan, "Skin cancer multiclass classification through different types of Convolutional Neural Networks," *2021 2nd International Conference on Computational Methods in Science & Technology (ICCMST),* 2021, pp. 84-87, doi: 10.1109/ICCMST54943.2021.00028.

[49] Ratul AR, Hamed MM, Lee W-S, Parimbelli E, “Skin lesions classification using deep learning based on dilated convolution.” *bioRxiv*, 860700, Dec. 2019. [online] Available: https://doi.org/10.1101/860700 .
[50] N. Abuared, A. Panthakkan, M. Al-Saad, S. A. Amin, W. Mansoor, "Skin Cancer Classification Model Based on VGG 19 and Transfer Learning," *2020 3rd International Conference on Signal Processing and Information Security (ICSPIS),* 2020, pp. 1-4, doi: 10.1109/ICSPIS51252.2020.9340143.
[51] Zhu CY, et al., “A Deep Learning Based Framework for Diagnosing Multiple Skin Diseases in a Clinical Environment,” *Front Med (Lausanne),* 16; 8:626369, Apr. 2021, doi: 10.3389/fmed.2021.626369.
[52] Wang S-Q, et al., “Deep learning-based, computer-aided classifier developed with dermoscopic images shows comparable performance to 164 dermatologists in cutaneous disease diagnosis in the Chinese population,” *Chin Med J.* ,133:2027–36, Sep. 2020, doi: 10.1097/CM9.0000000000001023.
[53] Fujisawa Y, et al., “Deep learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis,” *Br J Dermatol,* 180:373–81, Feb. 2019, doi: 10.1111/bjd.16924. Epub 2018 Sep 19.
[54] Liu Y, et al., “A deep learning system for differential diagnosis of skin diseases,” *Nat Med. 26*, pp. 900–908, 2020. [online] Available: https://doi.org/10.1038/s41591-020-0842-3
[55] R. Maiti, P. Agarwal, R. R. Kumar and A. Bhat, "Detection Of Skin Cancer Using Neural Architecture Search with Model Quantization," *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2021, pp. 1807-1814, doi: 10.1109/ICICCS51141.2021.9432190.
[56] Adegun, A., Viriri, S., “Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art,” *Artif Intell Rev 54*, 811–841, Feb. 2021, DOI:10.1007/s10462-020-09865-y.
[57] Khan, A., et al. “A survey of the recent architectures of deep convolutional neural networks,” *Artif Intell Rev 53*, 5455–5516, 2020. [online] Available: https://doi.org/10.1007/s10462-020-09825-6
[58] Zhu S, Zhu C, Wang W., “A New Image Encryption Algorithm Based on Chaos and Secure Hash SHA-256,” *Entropy (Basel).* 19;20(9):716, Sep. 2018, doi: 10.3390/e20090716.
[59] “Melanoma Detection Dataset”, Kaggle, https://www.kaggle.com/datasets/wanderdust/skin-lesion-analysis-toward-melanoma-detection (accessed Nov. 14, 2022)
[60] Hajian-Tilaki K., “Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation,” *Caspian J Intern Med*., 4(2):627-35, 2013. [online] Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3755824/
[61] Stehman, Stephen V., "Selecting and interpreting measures of thematic classification accuracy," *Remote Sensing of Environment*, 62 (1), pp. 77–89, 1997. [online] Available: https://doi.org/10.1016/S0034-4257(97)00083-7
[62] “Confusion Matrix”, Wikipedia, https://en.wikipedia.org/wiki/Confusion_matrix (accessed Nov. 14, 2022)

Acknowledgement

Special thanks to Prof. Sebastian Link for showing me the direction of this project. I would like to thank Mr. Yaniv Gal from Kahu AI for offering me this opportunity to do this project. I am grateful for the help of Mr. Kerry Carlyle and Mr. Thomas Rademaker in the company. For the last part of project, I would like to thank Prof. Patrice Delmas for granting me the access to IVS lab that I could use GPU in the lab to finish the AI simulation part of the work.

Thank my family for supporting me!

Appendix A
pseudo-code

Algorithm 1, Tracking the modification of AI database

```
Algorithm 1 : Data split and comparison

Input:  df1: old_status_image_db.csv
        df2: new_status_image_db.csv

Datasplit:
        split col[name] into ['photo'],['L0'], ['L1'], ['L2'],['data_type']
        output: splitted df1,df2
Data compare:
        df1 outerjoin df2 on ['name']
        df1_left_only <- df_delete # deleted records
        df2_right_only <- df_extra #extra records including new and label change
        df_extra[['name'] not in df1['name']] -> df_new # new images added
        # label changed L0,L0 & L1,L0 & L1 & L2,L1,L1 & L2,L2,L0 & L2
        # L0 change
        df_L0 <- df_extra[['photo','L1','L2'] in df1['photo','L1','L2']]
        #L0 & L1 change
        df_L0L1 <- df_extra[['photo','L2'] in df1['photo','L2'] AND ['L0,L1'] not in df1['L0,L1']]
        # L0 & L1 & L2 change
        df_L0L1L2 <- df_extra[['photo'] in df1['photo'] AND ['L0,L1,L2'] not in df1['L0,L1,L2']]
        # L1 change
        df_L1 <- df_extra[['photo','L0','L2'] in df1['photo','L0','L2']]
        # L1 & L2 change
        df_L1L2 <- df_extra[['photo','L0'] in df1['photo','L0'] AND ['L1,L2'] not in df1['L1,L2']]
        # L2 change
        df_L2 <- df_extra[['photo','L0','L1'] in df1['photo','L0','L1']]
        # L0 & L2 change
        df_L0L2 <- df_extra[['photo','L1'] in df1['photo','L1'] AND ['L0,L2'] not in df1['L0,L2']]

        result <- concat(df_delete,df_new,df_L0,df_L0L1,df_L0,df_L0L1,df_L0L1L2,df_L1,
                  df_L1L2, df_L2,df_L0L2)
Database upload:
        SQL connection settings
        upload(result)
        update logfile with username, appname, modification
```

Algorithm 2, Image-hashing in AI database.

```
Algorithm 2 : Image hashing and error tracking
Input:  root_uri of AI database

Imagehashing:
        for path_root in root_uri.iterdir() do
                for path, subdirs, filenames in path_root do
                        if file match pattern(.jpg)
                                full_path <- join (path,filename)
                                with open(full_path) as f do
                                        bytes <- f.read()
                                        hashcode <- hashlib.sha256(bytes)
                                        output_imagehash <- list('name','path','hashcode')
                        end if
                end for
        end for

Database upload:
        SQL connection settings
        upload(output_imagehash) as table imagehash'n'
Error tracking:
        SQL_query( imagehash'n'['hashcode'] = imagehash'n-1'['hashcode'] AND
                   imagehash'n'['name'] <> imagehash'n-1'['name']
        output SQL_query result as error.csv
Logfile update:
        update logfile with username, appname, No. of errors, error['name']
```

Algorithm 3, Result analysis for tracking the AI performance.

```
Algorithm 3 : AI result analysis

#column_names name(full_path + photo name), L0, L1, L2 labels with their AI scores
# path represents the actual labels of each image file
Input:  AI_result.csv


Namesplit: # to get the actual label of each image file
        split col[name] into L0, L1, L2 actual labels
        df['L0_actual'] <- convert L0 labels to two classes 0 and 1
        df['L1_actual'] <- convert L1 labels to four classes 0, 1, 2, 3
        df['L2_actual'] <- convert L2 labels to nine classes 1 ~ 9
        # group total 39 labels of L2 class to get sum of AI scores for each 9 subclasses
        if 'pattern' in col_names in df:
                df['L2_label'] <- df[col_names containing 'pattern'].sum_per_row
        end if

Binary_classification: # for two classes classification,
                       # eg. benign vs malignant, melanoma vs. non-melanoma
        FPR, TPR, Threshold <- ROC_curve(df['actual_label'], df['predicted_label']
        best_threshold <- threshold[tpr >=0.95]
        # output predicted class
        df[class_predicted] <- (0 if df['predicted_label'] < threshold else 1)
        output : df['L0_predicted'], df['L1_melanoma_predicted'],df['L2_melanoma_predicted']
L1_classification: # for four classes L1 classification
        df['L1_predicted'] = argmax(df['L1_columns']) # return indices of maximum values
        df['L1_predicted][where df['L1_melanoma_predicted']=1, change class to melanoma]
        output:  df['L1_predicted']
L2_classification: # for nine classes L2 classification
        df['L2_predicted'] = argmax(df['L2_columns']) # return indices of maximum values
        df['L2_predicted][where df['L2_melanoma_predicted']=1, change class to melanoma]
        output: df['L2_predicted']

Performance_evaluation:
        compute TP,FP,TN,FN,TPR,TNR,1/PPV, NPV, FPR, FNR, Accuracy, PT
Plot_confusion_matrix: # including normalized and non-normalized confusion matrix
        L0 classification
        L1 melanoma vs. non-melanoma
        L2 melanoma vs. non-melanoma
        L1 classification
        L2 classification
```