Saliency-Enhanced Content-Based Image Retrieval in Dermatology Imaging

MASTER THESIS

Department of Physics
ETH Zürich
Radio Oncology
University Hospital Zürich (USZ)
May 5, 2022

Mathias Gassner
mgassner@student.ethz.ch

SUPERVISION

Dr. Andreas Adelmann
andreaad@ethz.ch

Dr. Javier Barranco Garcia
javier.barrancogarcia@usz.ch

PD Dr. Stephanie Tanadini-Lang
stephanie.tanadini-lang@usz.ch

Prof. Dr. med. Ralph
ralph.braun@usz.ch

ETH zürich

USZ Universitäts Spital Zürich

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Saliency-Enhanced Content-Based Image Retrieval for Dermatology Imaging

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
Gassner	Mathias Josef

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
Zürich, 08. April 2022	

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.

Contents

GitHub repository for reproducing the evaluation results.

Abstract

Skin cancer rates are increasing around the world. Artificial intelligence (AI) tools have shown remarkable results in the research area of dermatology imaging and will play a significant role in tackling the increasing number of patients and cancer cases. Content-based image retrieval (CBIR) based on deep convolutional neural networks aims to retrieve the most similar images from a large-scale dataset against a query image. This tool has a wide acceptance among clinicians and can help to augment their diagnosis by displaying similar, previously diagnosed images. Usually, the relevant information of skin lesion images is spatially restricted. In this thesis, we propose a novel CBIR algorithm that increases the focus on the seemingly relevant regions of the skin lesion images by using saliency maps. We have evaluated our model against a conventional CBIR model based on deep convolutional neural networks and have observed a significant improvement in the quantitative retrieval. To support the relevancy of the retrieved images, we developed an online survey for dermatologists and residents of the University Hospital Zurich. By supporting the participants with six retrieved images and their labels, the diagnosis accuracy of diagnosing skin lesion images increased by 22% from the evaluation with no additional help.

List of Abbreviations

AI Artificial Intelligence

CBIR Content-Based Image Retrieval

CNN Convolutional neural network

DL Deep learning

IG-CBIR Interpretability-Guided Content-Based Image Retrieval

ISIC International Skin Image Collaboration

ML Machine learning

NN Neural network

RNN Recurrent neural network

SE-CBIR Saliency-Enhanced Content-Based Image Retrieval

SIIM Society for Imaging Informatics in Medicine

1 Introduction

Skin cancer rates are increasing around the world. Unfortunately, this applies also to melanoma, its most deadly type. Among Caucasians, melanoma rates increase by 3-7% every year [1]. Even though the mortality rates have stabilized, probably reflecting the efficacy of new systematic treatments, the demand for expert consultations is increasing. Adding artificial intelligence models to the workflow of clinicians could improve their efficiency and accuracy in skin lesion diagnosis.
Even though, several algorithms have been outperforming dermatologists in the classification of skin lesion images in recent years, the transition to real-world consultation has happened just in a few exceptions [2] [3]. Another tool to augment the clinical assessment of skin lesions are content-based image retrieval (CBIR) systems. CBIR systems are tools to search a large-scale dataset for the most "similar" images against a query image. CBIR systems have been described as one of the most promising tools for applying AI in the medical field and, additionally, being widely accepted among clinicians [4], [5], [6]. This thesis proposes a novel CBIR algorithm for skin lesion images.

1.1 Thesis Outline

The theoretical backbone of the project is developed in section 1. General information about skin cancer and a short introduction of artificial intelligence (AI) in the area of medical images is given. AI and its challenges in the field of dermoscopic data are analyzed more thoroughly before finishing this chapter with related work.
Section 2 aims to introduce the dataset used for our study, the HAM10000 (Human Against Machine with 10000 skin lesion images) dataset.
A more detailed explanation of the AI methods and techniques used in our proposed algorithm, together with an introduction of the algorithm itself, is found in section 3. Furthermore, the training and the evaluation methods are described here.
In section 4, the results of the quantitative and qualitative evaluation of the image retrieval are presented and analyzed.
Finally in section 5, we summarize the main results presented in the past sections and provide an outlook for feasible future projects.

1.2 Skin Cancer

Skin cancer is one of the fastest increasing forms of cancer. With sun exposure being its leading cause, the incidence is highly dependent on the solar exposure of the area and the pigmentation of the people's skin. Among Caucasians, 35-45% of all neoplasms account to skin cancer [7], where UV radiation is the cause in 90-95% of all skin cancer cases [8]. Among non-Caucasians, effective pigment protection leads to relatively rare cases of skin cancer, accounting to 1-2% of all neoplasms in the black population, to 2-4% among Asians, and 4-5% among Hispanics [9].
As a quick side-note before going deeper into the topic of skin cancer, lesions with a non-cancerous skin-growth are called *benign*, whereas they are referred to as *malignant* if the growth is cancerous.
The American Society of Clinical Oncology divides skin cancer into four main types [10]:

- The most common skin cancer is *basal cell carcinoma (BCC)*. Approximately four in five non-melanoma skin cancers develop from this cell. This type is primarily found in the neck and head area and is mainly caused by UV radiation or radiation therapy as children. BCC usually grows slowly and rarely spreads to other parts of the body.

- Most of the epidermis, the outermost skin layer, is made up of flat, scale-like cells called squamous cells. Around 20% of non-melanoma skin cancers develop from these cells, and these cancers are called *squamous cell carcinomas*. Sun exposure is again the main cause, and this type of cancer can be found across all regions of the skin. About 2% to 5 % of squamous cell carcinomas spread to other body parts.

- *Merkel cell cancer* is a highly aggressive or fast-growing, rare cancer. It starts in hormone-producing cells just beneath the skin and in the hair follicles. Merkel cell cancer may also be called neuroendocrine carcinoma of the skin and is mainly found in the head and neck region.

- There are scattered cells called melanocytes where the epidermis meets the skin's next layer, the dermis. These cells produce the pigment melanin, which gives skin its color. *Melanoma* starts in melanocytes and is the most serious type of skin cancer. Even though it accounts for only about 1% of all skin cancers, it causes around 75% of the deaths from skin cancer.

Other rare and less severe types of skin cancers are *cutaneous lymphomas*, *Kaposi sarcoma*, *skin adnexal tumors*, and *sarcomas*.
Focusing on melanoma, the most deadly type of skin cancer, SEER (Surveillance, Epidemiology, and End Results) stages can be used to categorize melanoma according to cancer growth. The stages are: (1) Localized: no sign the cancer has spread beyond the skin where it started; (2) Regional: The cancer has spread beyond the skin where it started to nearby structures or lymph nodes; and (3) Distant: the cancer has spread to distant parts of the body, such as the lungs, liver, or skin on other parts of the body. The stage-specific 5-year survival rate of American patients diagnosed with melanoma skin cancer between 2011 and 2017 are: (1) Localized: 99%; (2) Regional: 68%; and (3) Distant: 30% [11].
Therefore, early diagnosis of melanoma is important. Without any doubt, AI will play a major role in improving the diagnosis of skin cancer. In recent years, the research in this area has been rapidly growing, more and more data has been acquired and promising first applications have

been developed, which persuaded an increasing number of dermatologists to use AI on a daily basis. Before going into more detail on the challenges of applying AI for skin cancer diagnosis, AI is introduced more generally with a focus on medical image classification subsequently.

1.3 AI/ML introduction with a focus on Medical Image Classification

Artificial intelligence (AI) and machine learning (ML) have proven to be very efficient tools in many research areas of medical imaging [12], [13]. In this subsection, we give a brief introduction to AI and ML targeting the area of medical image classification.
Many definitions of AI exist, dating back to 1950, when Alan Turing introduced the Turing test as the first definition of AI, calling it computational intelligence at this point [14]. Since then, several changes and adaptations have been made. The leading AI textbooks have somewhat similar definitions, which Shane Legg and Marcus Hutter grouped in 2007, and defined AI as:

> *"Intelligence measures an agent's ability to achieve goals in a wide range of environments."*

Features such as the ability to learn and adapt, or to understand, are implicit in the above definition as these capacities enable an agent to succeed in a wide range of environments [15]. Due to the different definitions and knowledge of the audience, the handling of this term varies in the literature. Trying to be in conformity with the literature, in this thesis, AI is used when referring to general or medical topics, where machine learning (ML), a subset of AI, and deep learning (DL), a machine learning technique, in parts describing the techniques and algorithms. Subsequently, ML with its techniques, methods, terminology, and features is shortly introduced. Machine learning refers to algorithms that infer information from data in an implicit way [16]. Figure 1.1 shows a task-oriented overview of ML, including many of the well-established methods. Since our aim is to classify medical images, we focus on supervised learning using neural networks in the form of convolutional neural networks (CNNs). A brief description of these terms verifies this choice.
The following of this subsection is based on the lecture notes "Introduction to Machine Learning for the Sciences" of Neupert et al. [16]. We used their definitions and adapted some of their descriptions for our purpose.
Supervised learning is the term for a machine learning task, where given a dataset consisting of input-output pairs $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)\}$, the task is to "learn" a function which maps input to output $f : \mathbf{x} \rightarrow y$. The output data is called the *ground truth* and is sometimes also referred to as *labels* of the input. For example, input-output pairs consisting of arm X-ray images labeled either with "broken bone" or "healthy" may be used. *Unsupervised learning* tasks rely exclusively on the input data with the goal of either "learning" the underlying structure of the data or generating new data. On the other hand, *reinforcement learning* does not fall in a data-driven category. The task of a reinforcement learning agent is to interact with an environment through actions, which on the one hand, change the state of the agent, and on the other hand, lead to a reward. Here, the goal is to maximize the reward. Resuming to supervised learning, there are two types of tasks: *Classification* and *Regression*. In a regression problem, the output y is a continuous number or vector, whereas in a classification task, y is a discrete variable corresponding to a classification category. With the advent of deep learning along with faster

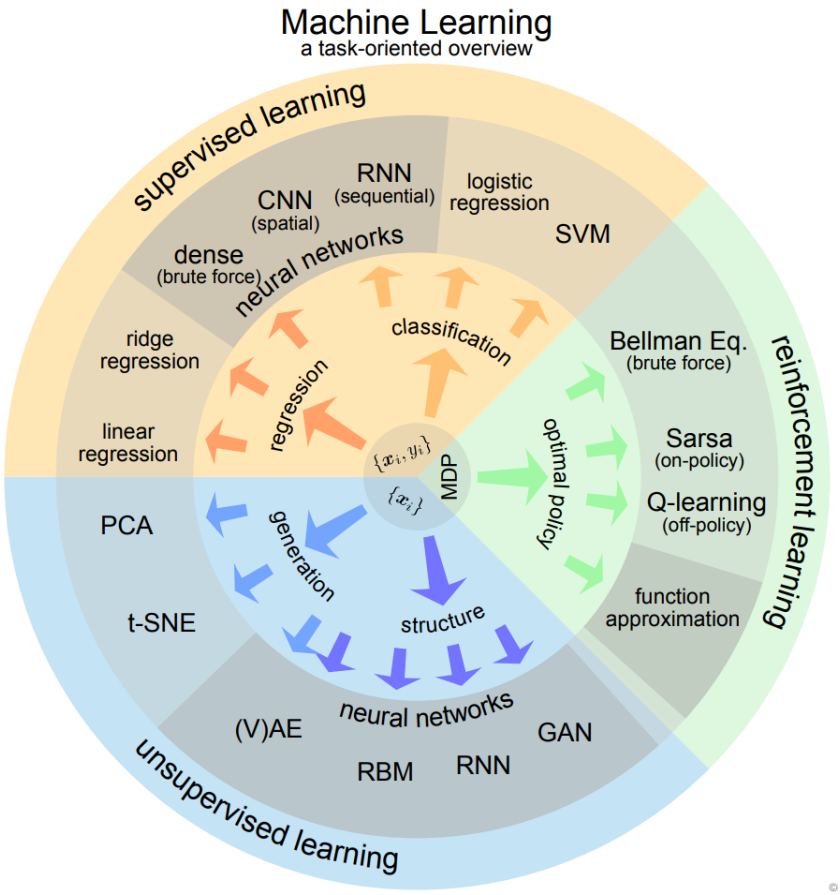


Figure 1.1 – Machine Learning overview [16].

hardware, and rapid increase in data and its quality, neural networks outperform traditional ML methods in most tasks, especially if enough data is available. There is a thorough introduction to deep learning and neural networks in the subsequent paragraph. For the time being, neural networks consist of layers, and the term deep learning refers to using neural networks with more than three layers. The two main neural networks for classification tasks are *convolutional neural networks* (CNNs) and *recurrent neural networks* (RNNs). Where RNNs are used for sequential data, such as time series, CNNs perform well on data where geometric information is essential, such as images.
In order to interpret the results and detect possible issues of image classification using CNNs, however, a more thorough understanding of neural networks is essential.
A *neural network* (NN) is a machine learning algorithm designed as an analogy to how biological organisms process information. Biological brains contain neurons, electrically activated nerve cells, connected by synapses that facilitate information transfer between neurons. The machine learning equivalent of this structure, the so-called artificial neural networks or neural networks in short, is a mathematical function developed with the same principles in mind. The composition of a simple (feed-forward) neural network is shown in Figure 1.2. The dots are the neurons, which are mathematically represented with activation functions. Vertically aligned neurons form a layer, and information is passed from left to right by the connection lines. The input and output layers (blue and violet, respectively) are called visible layers, as they are directly

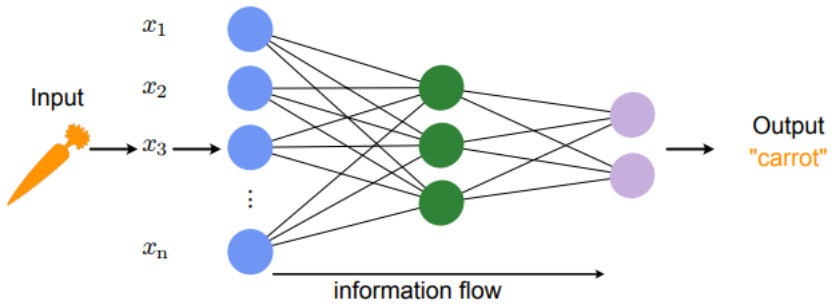


Figure 1.2 – Neural Network. Graphical representation and basic architecture [16].

accessed. All the other layers in between them are neither accessible for input nor provide any direct output, and thus are called hidden layers. In this example, we have a single hidden layer (green), and since all the possible neuron connections to the previous and subsequent layers are made, it is called a fully connected (FC) layer. If there are at least two hidden layers present, the network is referred to as a *deep neural network*, and the technique of employing such a network is called *deep learning* (DL). Since this is mainly the case, the terms are often used interchangeably.
Each neural network can be written by means of a mathematical functional $y^* = F[\mathbf{x}]$. In the case of Figure 1.2, the output y^* (e.g. "carrot") is the suggested label to the input $\mathbf{x}$ (e.g. image of carrot). Without going into too much detail, the functional is created by the architecture (e.g., how many layers, which neurons are connected) of the network and the activation functions of the neurons. Each activation function is dependent on the bias b of the neuron, and the weights W representing the "strength" of each connection line to this neuron. Therefore, the functional $F[\mathbf{x}]$ is parameterized by the weights and biases $\theta = \{W, b\}$.
The goal of *training* a neural network is to adapt the weights and biases in a way that $F[\mathbf{x}]$ gets close to $\mathbf{y}$, where $\mathbf{x}$ and $\mathbf{y}$ represent all inputs and ground truths of a dataset $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$. This task is achieved by minimizing a loss function $\mathcal{L}(F[\mathbf{x}]; y) = \mathcal{L}(\theta)$ parametrized by θ, the weights and biases. Since $\mathcal{L}$ is typically a high-dimensional function, the minimum is analytically intractable. Therefore, an iterative optimization algorithm combined with *backpropagation* is used to step-wise adapt the parameters in order to "travel" closer to a minimum of $\mathcal{L}$. Backpropagation profits from the chain rule of differentiation to efficiently calculate the derivative of $\mathcal{L}$ with respect to all parameters using parallel computing. With the loss functions derivative, the parameters can be adjusted using an optimization algorithm. *Gradient descent*, for example, adapts the parameters in each optimization step by

$$\theta_\alpha \to \theta_\alpha - \eta \frac{\partial \mathcal{L}(\theta)}{\partial \theta_\alpha}. \tag{1.1}$$

Here, θ represents the set of all weights (W) and biases (b) with θ_α being an arbitrary choice of this set. The *learning rate* η specifies the step size of the optimization procedure.
Often, not all of the training data is provided to the network at once, since it is computationally costly and sometimes does not lead to the best performances. Therefore, the input data is divided into so-called *batches*, a group of training data that is fed into the network together.

Then, one *epoch*, which describes a training cycle, is given by:

1. Dividing training data into batches: *batch 1*, ..., *batch N*
2. For $i = 1$ to N:
   (a) feed batch i to network
   (b) calculate loss $\mathcal{L}(\theta)$
   (c) update parameters θ with respect to the optimization algorithm

   This procedure is shown in Figure 1.3.

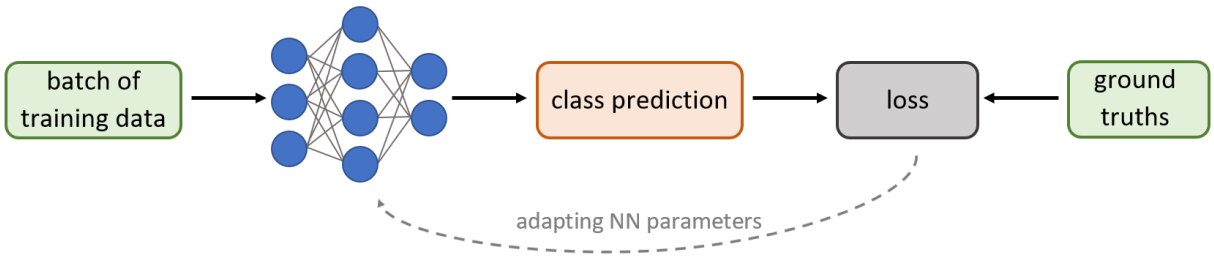


Figure 1.3 – Single optimization step by feeding one batch of data to the neural network. Green boxes represent the training data (input and ground truths); the loss (grey) is calculated from the NN's class predictions (orange) and the ground truths; parameters are adapted according to a loss-dependent optimization algorithm.

Often, the training data is split into a validation and training set, where the validation set is not used for training but for evaluation to see whether the training is going in the right direction. As a quick side note, before continuing with CNNs, the learning rate η and the size of the batches have to be set before training, and thus are called *hyperparameters*. Finding a good set of hyperparameters is one of the most challenging parts in deep learning. The aim is to find a set of parameters to achieve good performance without *overfitting*. Overfitting means that the network has learned specificities of the dataset it was presented with, rather than the abstract features to fulfill the task well on new data.
Neural networks containing *convolutional layers* are called *convolutional neural networks* (CNNs). The key idea behind the convolutional layers is to identify certain local patterns in the data. Each convolutional layer consists of a set of learnable *filters* (or *kernels*). Unlike the general neural network introduced earlier, in CNNs the filters are trained to learn the important patterns of the data. In Figure 1.4, we see an example of applying a CNN for neurosurgical images [17]. Part A on its left side shows the architecture of the network: an input layer, five convolutional layers, followed by two fully connected layers, and an output. In the training process of such a CNN, the filters of the convolutional layers are trained parallel to the weights and biases of the two fully connected layers. Part B visualizes the filters of the first convolutional layer. Generally, filters in the first convolutional layers are sensitive to local patterns, while ones in the later layers recognize larger structures. This is due to pooling layers in between the convolutional layers grouping information together, leading to a zoom-out effect.
The following subsection focuses in more detail on possible AI implementations in dermatology and its challenges, such as explainability and generalization.

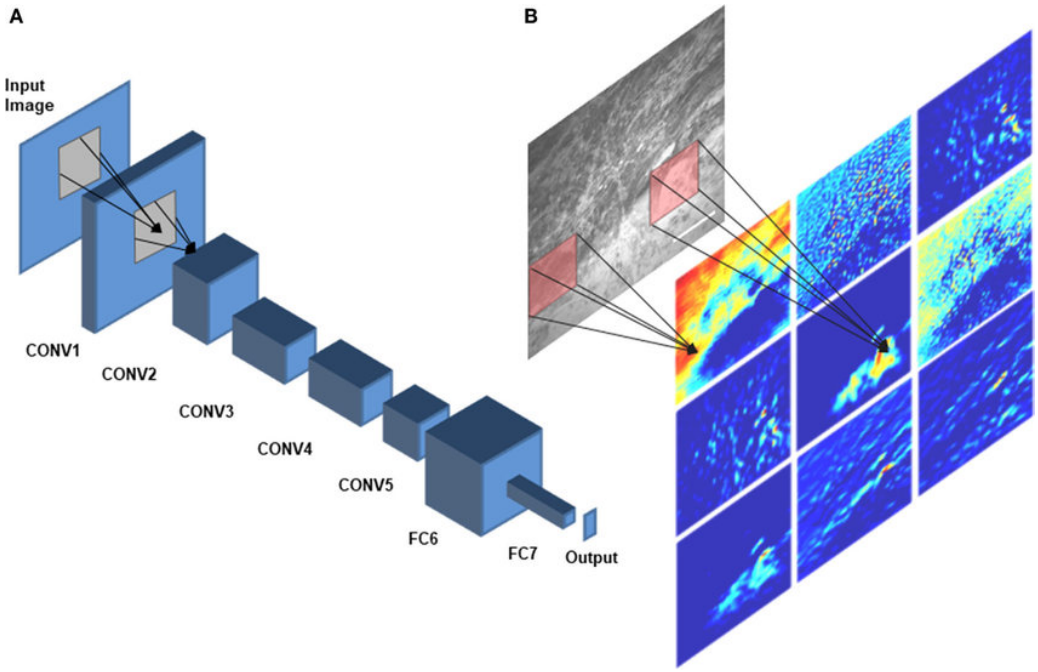


Figure 1.4 – A: Example of a convolutional neural network; B: Visualization of different filters of the first convolutional layer [17]

1.4 AI in Dermatology Imaging

There are many applications of artificial intelligence in dermatology imaging, such as teledermatology or telepathology. The algorithm developed in this thesis can be used to augment the clinical assessment in face-to-face consultations. Thus, this part focuses solely on this application area of AI algorithms in dermatology imaging.
With the emergence of a yearly AI for dermatology imaging challenge hosted by the International Skin Image Collaboration (ISIC) in 2016, several AI models have outperformed board-certified dermatologists in the classification of dermatology images of skin lesions from 2018 on. Therefore, applying AI, especially for less experienced dermatologists, who have not been exposed to thousands of patients over many years in order to make confident and accurate diagnoses, should undoubtedly improve their decision-making. Nevertheless, a recent study has shown that only 5% of dermatologists use AI on a daily basis, and 85% have not used it at all [3]. The subsequent paragraphs dive into the (still present) challenges of applying AI to augment the clinical assessment, trying to explain why it hasn't taken off yet.

Data: more than images

It is important to point out that the performance comparisons from earlier, stating that AI beats dermatologists in skin lesion classification, were made in an ideal setting for the AI by taking just dermatology images into account and executing the evaluation on data reflecting on the training data of the AI models. Furthermore, in a face-to-face consultation, dermatologists may consider much more information, rather than just the visual appearance of a lesion, to decide whether a suspicious lesion is benign, or if a follow-up appointment or biopsy is necessary. First of all, the patient-level contextual information highly influences the decision of dermatol-

ogists on how to treat a suspicious lesion. For example, a patient having a single skin lesion with melanoma characteristics is more likely to be diagnosed histopathologically than a patient covered with many of such lesions. The patient's background information, so-called metadata, such as gender, ethnicity, age, and body location of the captured skin lesion, can also alter the decision. Besides that, dermatologists take the patient's history, such as captured images of suspicious lesions from previous consultations, previous cancer cases, and family history of cancer, into consideration as well.
Adding these types of information to algorithms and datasets is an active field of research. Li and Zhuang have successfully added metadata to several image classification algorithms and observed superior performance compared to the algorithms without metadata fusion [18]. The dataset for the "SIIM-ISIC Melanoma Classification" challenge of 2020 centered their dataset to patient-level contextual data [19]. The intention of this dataset, consisting of 33,126 dermoscopic images of 2,056 different patients, is to develop AI models reflecting clinical scenarios more closely.

AI-challenges: Data, Generalization, Explainability

The data used for training should reflect the patients on whom the AI model will be applied. Ideally, all genders, ethnicities, cancer types, body locations, types of hair covers, et cetera, are included in the training set since the diversity of patients and their skin lesions is most likely not known in advance. It is impossible to fully remove the bias of the data with respect to the patient audience it will be used for due to a lack of available data and different underlying structures. Minimizing it before training an AI model is good practice. Nevertheless, some bias will remain.
In dermatology imaging, especially, there is underlying structure and bias in the data, which cannot be simply removed. To begin, most images in the common skin lesion datasets are from clinical assessments, where mainly suspicious benign lesions or hard-to-detect malignant lesions are recorded. The lack of clear cases in these datasets adds bias. Moreover, there are several ways to annotate the ground truth in dermoscopic images. Next to histopathology, labels are also assigned by follow-up appointments, expert consensus, or confirmation by in-vivo confocal microscopy. Of those, histopathology is the most accurate one but is mainly done for uncertain clinical diagnoses, especially for possible melanoma. Expert consensus and follow-up annotation can lead to a systematic error. This underlying structure could then be learned by the AI model. Hekler et al. showed in their study that AI classification algorithms trained data with biopsy-verified ground truths perform worse on test sets with dermatological annotated images than on biopsy-verified ones and vice-versa [20].
Ideally, AI should be developed in a *generalizable* fashion, meaning to perform well out of the closed-loop scenario, where training and testing are done on the same dataset. Even though deep neural networks have achieved remarkable performances in many tasks, they are prone to overfitting if trained with a limited amount of data. Since overfitting directly leads to a drop in generalization, *data augmentation* plays an important role in fighting overfitting. Data augmentation is a technique to enhance the amount of training data artificially, e.g., by rotating, mirroring, and cropping images. In image classification, augmentations help the network to "learn" the important features and help to prevent it from learning the datasets underlying structures which shouldn't affect the network's decisions. In the classification of skin lesions, the acquisition of images could add such underlying structures caused by marker ink, dark

corners, length scales, or gel bubbles. But augmentation comes with a cost: it decelerates training due to a larger amount of data and the loss of some information, e.g., by cropping or by adding noise. To give a negative example of generalization, Han et al. released their AI model trained on more than 20,000 images with a reported area under the receiver operating characteristic (AUROC) score of 0.91 as a web application [21]. Navarette-Dechent et al. tested this model on 100 high-quality images of the ISIC dataset resulting in an AUROC score of 0.29 [22]. There are several possible reasons for this performance drop for different test sets, with overfitting and biased datasets being the main ones and standardization of input images being another important reason to mention. Capturing skin lesions with different techniques, at different intensity and lightning conditions will lead to a drop in performance. Regarding the biased datasets, Han et al. observed a significant performance drop in evaluating their AI model trained on a dataset acquired in Asia and evaluated on a dataset acquired in Europe containing mainly Caucasian samples [23].
In a real-world application, not all of these issues can be resolved. In order for the users of these AI models to be aware of possible issues, *explainability* (or interpretability) plays a major role, especially if the users are no experts in AI. In general, explainability or interpretability in AI refers to opening up the "black box" and adding information on how the AI model decides, to understand its results better. One explainability method, the extraction saliency maps, is introduced in Section 3.1. Especially in the medical field, the explainability of AI models is highly desired due to ethical and legal reasons. In the European Union, a "Right to Explanation" was added to the *General Data Protection Regulation* (GDPR) in 2016, being effective since 2018 (stated in Section 4, Article 22, Recital 71) [1].

Clinical implementation

In order to be used in clinical diagnostics, AI should be fluently added to the workflow as a support tool without increasing time or reducing capacity but adding reliability and/or accuracy to their diagnosis. Furthermore, a safe and ethical implementation is required.
One possible implementation of an AI model to augment a clinical assessment and being able to fulfill the stated requirements is shown in Figure 1.5. To refer to the previous subsection, the images taken in the examination need to be captured in a standardized way for the AI to perform well.

Support tools in Dermatology Imaging

To support clinicians in the complex task of interpreting skin lesion images, several support systems have been developed by researchers, often referred to as Computer-Aided Diagnosis (CAD) tools. With the advance of deep learning, classification tools for pigmented skin lesions were developed, performing comparably to experts in the field [2]. However, the lack of reliability and interpretability lessens their ability as support tools. Another option to augment the clinical assessment are so-called content-based image retrieval (CBIR) systems. Those systems aim to find "similar" (previously diagnosed) images from a large-scale database against a query image. Since the users themselves decide which of the retrieved images are of importance, these systems add more leeway compared to classifiers to augment their decision. This leads to a

[1] Article 22 of the EU GDPR (https://www.privacy-regulation.eu/en)

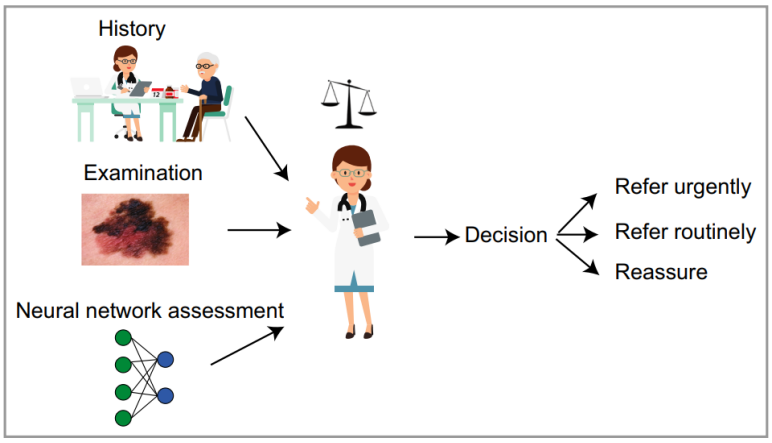


Figure 1.5 – Schematic showing hypothetical use of a machine learning algorithm to augment the clinical assessment [24].

wide acceptance of CBIR tools among clinicians [4]. Nevertheless, how these deep learning based CBIR systems decide which images are the most "similar" ones remains often unknown.

1.5 Related Work: Interpretability-Guided Content-Based Image Retrieval (IG-CBIR)

In 2020, Wilson Silva et al. published their work using a novel algorithm for *content-based image retrieval* (CBIR) of chest X-ray images [25]. Considering the fact that relevant information in medical images is typically spatially constricted, they developed a new architecture to enforce the algorithm focusing on the seemingly more important regions of each image, naming it *Interpretability-Guided Content-Based Image Retrieval* (IG-CBIR). The algorithm (see Figure 1.6) is trained in two steps: (1) a CNN model is trained to classify the possible disease located in the chest area; (2) *Saliency maps* are extracted from the classifier of step 1 and then used to fine-tune this classifier. The saliency maps highlight the pixels with respect to their contribution to the absolute classification of the CNN (see section 3.1 for a thorough explanation). Assuming that the CNN focuses on the clinically relevant areas, Silva et al. stated that fine-tuning the classifier with the extracted saliency maps enhances this focus due to the spatial bias of the maps, leading to an improvement of the CBIR.
Their study has shown an improvement in the quantitative and qualitative evaluation, respectively, of retrieving the most "similar" images from the fine-tuned classifier instead of the classifier trained in step 1.
It is essential to mention some points of concern in their evaluation. The used dataset already has a relatively small test set of 200 images, whereas its training set is more than 200,000. The authors reduced the test set further for their qualitative and quantitative evaluation. Due to the time limitations of radiologists, there was a reduction of the test by splitting it into query images and images to be retrieved for the qualitative evaluation, resulting in a rise in subjectivity. Nevertheless, executing the quantitative evaluation on the test set solely by splitting it similarly as in the qualitative evaluation was without reason. Taking all images of the test set as queries and retrieving from the 200,000 training images would lead to more objective results.

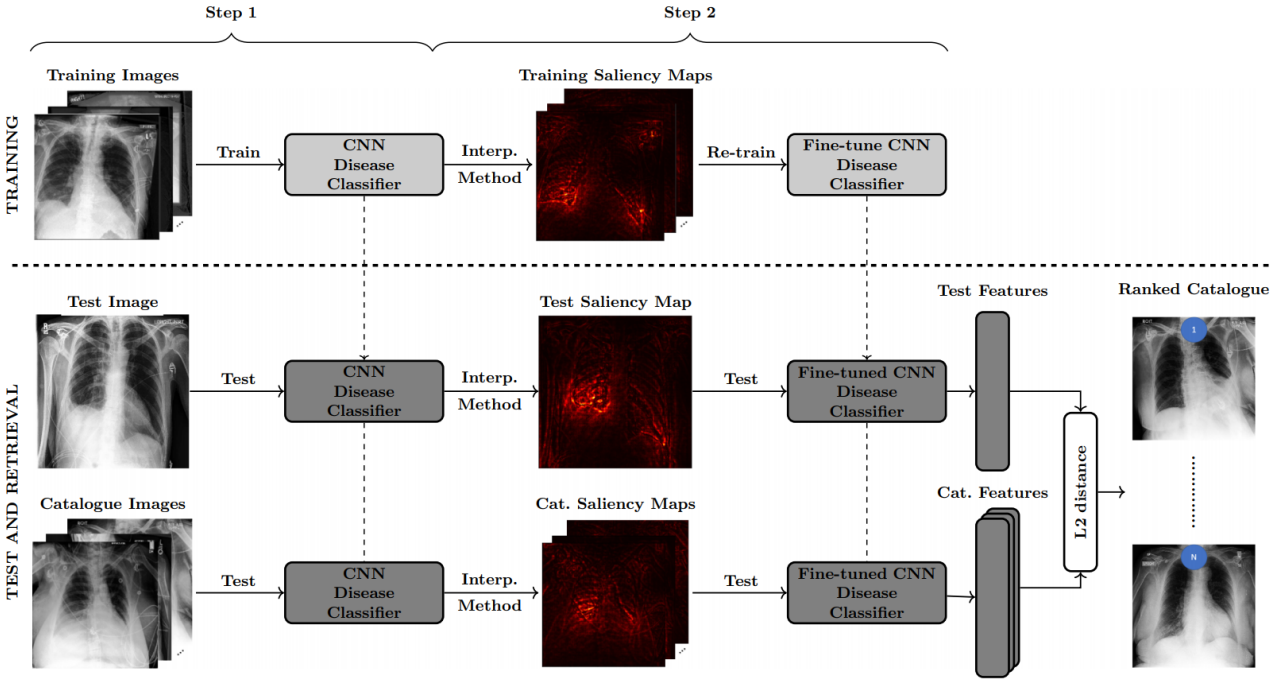


Figure 1.6 – Overview of the IG-CBIR algorithm [25].

2 Dataset

In this thesis, the HAM10000 ("Human Against Machine with 10000 training images") dataset by Tschandl, Rosendahl, and Kittler [26] was used. The authors collected dermoscopic images from different populations, acquired and stored by different modalities. With their goal of creating a high-quality dataset for machine learning tasks, a semi-automatic workflow (Figure 2.1) was created to organize and clean the data before saving it in a standardized format. The

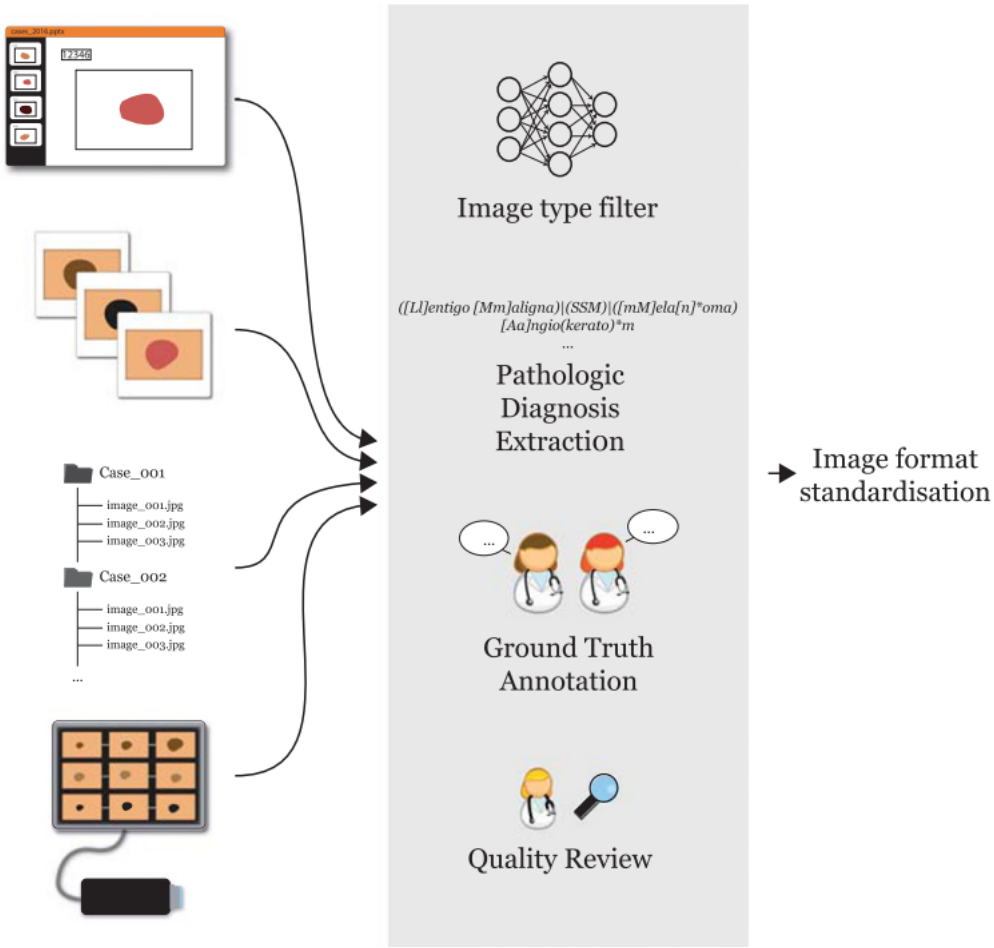


Figure 2.1 – Schematic workflow of dataset workup methods. Image and data content from different sources were entered into a pipeline to organize and clean data, with final images being standardized and stored in a common format [26].

importance of standardized images for AI applications has been pointed out in the introduction. Furthermore, the authors reviewed the ground truths of the collected images in order to reduce the dataset's bias, e.g., emerged from possible systematic errors in the ground truth labeling. They excised lesions with a histopathologic report but excluded cases with ambiguous diagnoses. Nevi lesions (introduced below) were only taken into account if they didn't show changes within three follow-up visits or in between two follow-ups separated by more than 1.5 years. Other lesions were verified using reflectance (or in-vivo) confocal microscopy, including a one-year follow-up. For lesions without a histopathologic report, microscopic verification, or follow-up, two experts verified the ground truth. Lesions, where no consensus of both experts' diagnoses

with previous ground truth was reached, were removed.
Low-quality images and insufficiently magnified images of small lesions were removed by the authors too. Those would decrease the model's performance but are also not related to a real-world scenario. Some of the lesions included in their dataset may appear more than once due to follow-up appointments or images taken at different magnifications or angles.
Their final dataset consists of 10015 skin lesion images with more than half, 53% to be precise, of lesions confirmed by pathology. The dataset includes cases of all important diagnostic categories in the realm of pigmented lesions. They assigned the lesions to seven classes with the following definitions

akiec
This class includes Actinic Keratoses (Solar Keratoses) and Intraepithelial Carcinoma (Bowen's disease). Both are common non-invasive variants of squamous cell carcinoma that can be treated locally without surgery.

bcc
Basal cell carcinoma is a common variant of epithelial skin cancer that rarely metastasizes but grows destructively if untreated. It causes the second most deaths from skin cancer behind melanoma.

bkl
"Benign keratosis" is a generic class that includes seborrheic keratoses ("senile war"), solar lentigo, and lichen-planus like keratoses (LPLK). Even though these three subgroups may look different dermatoscopically, they were grouped together due to their biological similarities and because they are often reported under the same generic term histopathologically.

df
Dermatofibroma is a benign skin lesion regarded as either a benign proliferation or an inflammatory reaction to a minimal trauma.

mel
Melanoma is a malignant neoplasm derived from melanocytes that may appear in different variants. Invasive and non-invasive (in-situ) melanomas are included, but non-pigmented, subungual, ocular or mucosal melanoma are excluded from this dataset. As stated in the introduction to skin cancer, melanomas correspond to three-quarters of all skin cancer deaths.

nv
Melanocytic nevi are benign neoplasms of melanocytes and appear in a myriad of variants, which are all included in this dataset. The variants may differ significantly from a dermatoscopic point of view. Since missing a melanoma can have severe consequences for the patient, nevi lesions with visual similarities to melanomas are captured more frequently, leading to an over-representation in this class.

vask
This class includes vascular skin lesions ranging from cherry angiomas to angiokeratomas and pyogenic granulomas, but also hemorrhages.

class	*akiec*	*bcc*	*bkl*	*df*	*mel*	*nv*	*vasc*	**total images**
# of images	327	514	1099	115	1113	6705	142	10015

Table 2.I – Summary of HAM10000 dataset.

Here, classes with benign lesions are colored green, whereas classes reflecting malignant lesions are colored red. The class *akiec* is colored orange because actinic keratoses and Bowen's disease are usually benign but develop skin cancer in about 10% of the cases [27], [28]. Since this dataset focuses only on pigmented skin lesions, Merkel cell carcinoma is not included.
The occurrence of each class in the dataset is presented in Table 2.I. The dataset is strongly imbalanced. Melanocytic nevi contribute to two-third of all skin lesion images, and the classes *akiec*, *bcc*, *df*, and *vasc* are each represented by five or less percent of all images. A dataset reflecting a real-world scenario would include even more nevi cases and far fewer melanomas. Since it is by far the most deadly skin cancer type, melanomas are captured with a much higher frequency than other skin lesion types during clinical assessments.

3 Methods and Algorithm

This section introduces our proposed *Saliency-Enhanced Content-Based Image Retrieval* (SE-CBIR) algorithm. Prior to this, the different methods and techniques applied in SE-CBIR, such as transfer learning, saliency maps, and content-based image retrieval (CBIR), are explained. Finally, the training procedure is explained, followed by a discussion of the different methods for evaluating image retrieval systems, including our choices and their substantiation.

3.1 Methods

3.1.1 Transfer Learning and Fine-Tuning

The general approach for training a CNN classifier on a dermoscopic dataset is by taking advantage of a network trained on much larger, generic image datasets, such as *ImageNet* [29] or *noisy student* [30]. Those so-called pre-trained networks can be seen as generic models of the visual world. By "transferring" knowledge of a pre-trained network to a different task, such as skin lesion classification, better performances can be achieved with less training time compared to training a neural network of the same architecture from scratch. To be more precise, "transferring" knowledge refers to initializing the network for the new task with the trained parameters of a generic model. This approach is thus called *transfer learning*. The transfer learning strategy was applied for both classifiers of our developed SE-CBIR algorithm that is introduced later in this chapter. Subsequently, this approach is explained in more detail for each classifier. Figure 3.1 shows our transfer learning approach for both classifiers. Here, the first row shows the generic network, the 3-channel CNN classifier is illustrated in the middle row and the 4-channel CNN classifier in the last row. We decided to call the approach for training the 3-channel classifier transfer learning but refer to it as *fine-tuning* for the 4-channel classifier. Fine-tuning is often used interchangeably with transfer learning but is mainly used when the networks and their tasks change just slightly, which is why we chose this term for the 4-channel model.
The model names, 3-channel and 4-channel CNN classifier, were given according to their input data types. The first classifier takes skin lesion images as an input, however, pairs of skin lesions and saliency maps are fed to the second network. Color images are read by the computer in three channels, e.g., RGB channels. Each pixel is encoded to three intensity values corresponding to the colors red, green, and blue. The pixels of grey-scale images, such as saliency maps, can be encoded with a single intensity value. Therefore, the combination of skin lesions and their saliency maps are fed to the network as 4-channel inputs.

Transfer learning (3-channel classifier)
For the implementation of a transfer learning approach, one needs to choose a pre-trained network in a first step, followed by changing the classification layers according to their specific task, such as the classification of skin lesions from the HAM10000 dataset in our case.
The *EfficientNets*, are large CNN classifiers with a novel scaling technique [31]. Trained on either the *ImageNet* or *noisy student* dataset, they have been the best performing networks with respect to the number of training parameters at their invention in 2019. Additionally, most well-performing teams of the most recent SIIM-ISIC dermoscopy challenge from 2020

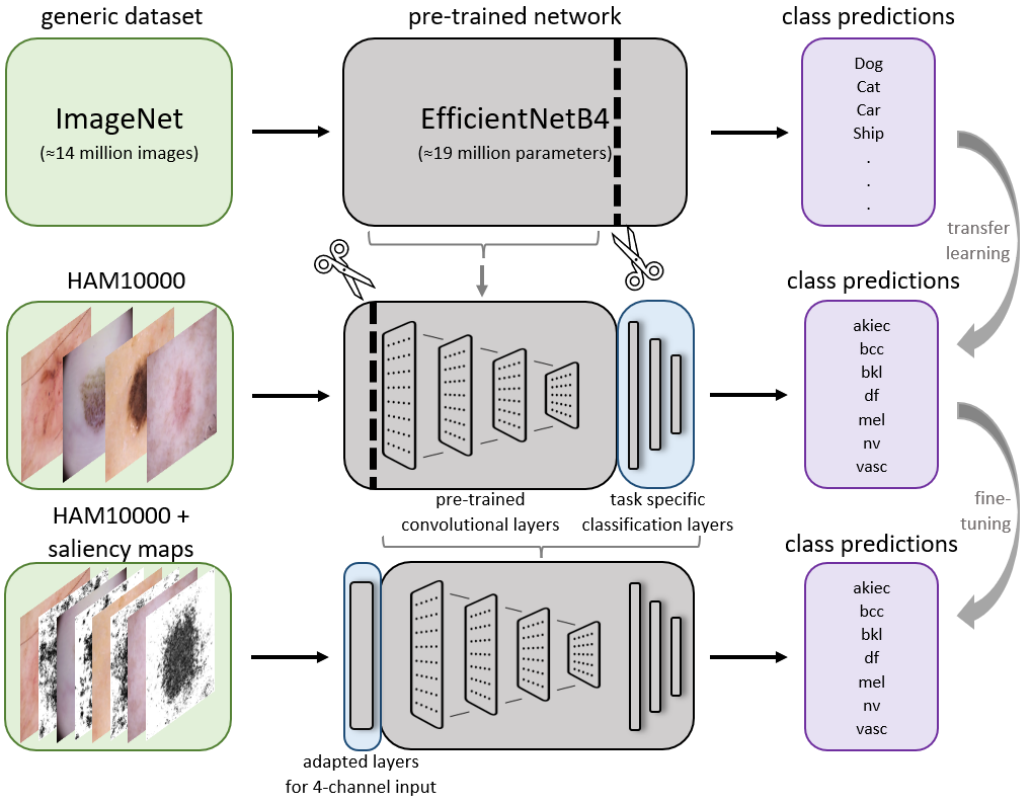


Figure 3.1 – The top row shows the pre-trained EfficientNetB4 on the ImageNet dataset. The middle row illustrates the transfer learning approach by adding task-specific classification layers to the previous architecture. The last row shows the changes to be made for fine-tuning this previous model on skin lesion-saliency map pairs of the HAM10000 dataset.

have used EfficientNets, such as the winning team [32]. Thus, we chose the EfficientNetB4, with 19 million parameters optimized for the ImageNet dataset with inputs of size 380x380 pixels. Furthermore, we replaced the EfficientNetB4's classification layers with our own, task-specific ones. The choice of the new layers is a rather heuristic one. We added a 2D average pooling layer and a batch normalization layer, followed by two combinations of a dropout and a fully-connected layer. For the dropout layers, we chose a dropout rate of 0.2 and dense layers consisting of 64 and 7 neurons, respectively.

Fine-Tuning (4-channel classifier)

As illustrated in Figure 3.1, we used the trained 3-channel CNN classifier for "transferring" its knowledge to the 4-channel model. Different from the previous transfer learning approach, the task of classifying skin lesions remained the same, but the input data type changed with the addition of saliency maps. Hence, no changes in the classification layers were necessary. The input layer and the EfficientNetB4's first convolutional layer, on the other hand, were adapted for the skin lesion-saliency map pairs, keeping the rest of the 3-channel CNN classifier model and its trained parameters.

3.1.2 Saliency Maps

With the emerging field of explainable AI, a variety of saliency methods have been developed, such as *vanilla gradient* [33], *SmoothGrad* [34], and *integrated gradients* [35]. Saliency methods aim to create a map highlighting the pixels which are relevant for the network's classification of a particular input image, a so-called *saliency map*. Due to the time limitations of this thesis and a large number of saliency map interpreters, we restricted ourselves to a gradient-only method, vanilla gradients, and a path-attribution method, integrated gradients. Gradient-only methods tell us how pixel-changes affect the prediction of a certain class. In contrast, path-attribution methods detect relevant pixels by summing over the gradients from a baseline input (e.g., grey image) to the particular image. Both methods were implemented using the *tf-explain* python package [36]. For our specific algorithm, we achieved better retrieval performances using the vanilla gradients to extract saliency maps. For this reason, we refer to vanilla gradient when using the term saliency map henceforth. Below, a more thorough explanation of saliency maps by the vanilla gradient method follows.
The saliency maps are derived by taking the gradient of the score function S_c with respect to the input image I_0

$$E_{grad}(I_0) = \frac{\delta S_c}{\delta I}\Big|_{I=I_0}. \tag{3.1}$$

This is motivated by the first-order Taylor expansion on the score function of the network's last layer

$$S_c(I) = w^T I + b \qquad \text{with} \qquad w^T = \frac{\delta S_c}{\delta I}\Big|_{I=I_0}. \tag{3.2}$$

Here, w are the weights and b the biases of this layer. The derivative of the score function is calculated similarly to the loss function's derivative during training. Both take advantage of the chain rule of differentiation, using backpropagation through the network. In more detail, a saliency map of a specific input image is extracted in three steps: (1) forward-pass of the input image through the CNN; (2) calculate the gradient of the class score with respect to the input image (Equation 3.1); and (3) visualize the gradients. In step (2), S_c is the class score of the class of interest. All other classes are set to zero. For skin lesions of the training set, we extracted the saliency map using the class score of the ground truth class, where we used the class score of the predicted class for images of the test set.
In Figure 3.2, three examples of skin lesions of the HAM10000 dataset (top row) and their corresponding saliency maps, extracted from the 3-channel CNN classifier, are shown. The intensity (0 = white, 255=black) of each pixel reflects the gradient of the class score. Therefore, the darker spots represent the relevant regions for the network's decision according to the vanilla gradient method. The saliency maps reflect primarily the lesion itself, of which some regions seem to be more relevant than others. Furthermore, sun damage (i.e., spots around the lesion of the center image) and, unfortunately, the corners of the right lesion are shown as relevant. Sun damage is indeed a vital characteristic in the diagnosis of skin cancer. Additionally, the saliency maps show little to no sign of the scale displayed next to the center lesion or the hairs of the right image, which both have no visual relevance for the diagnosis.

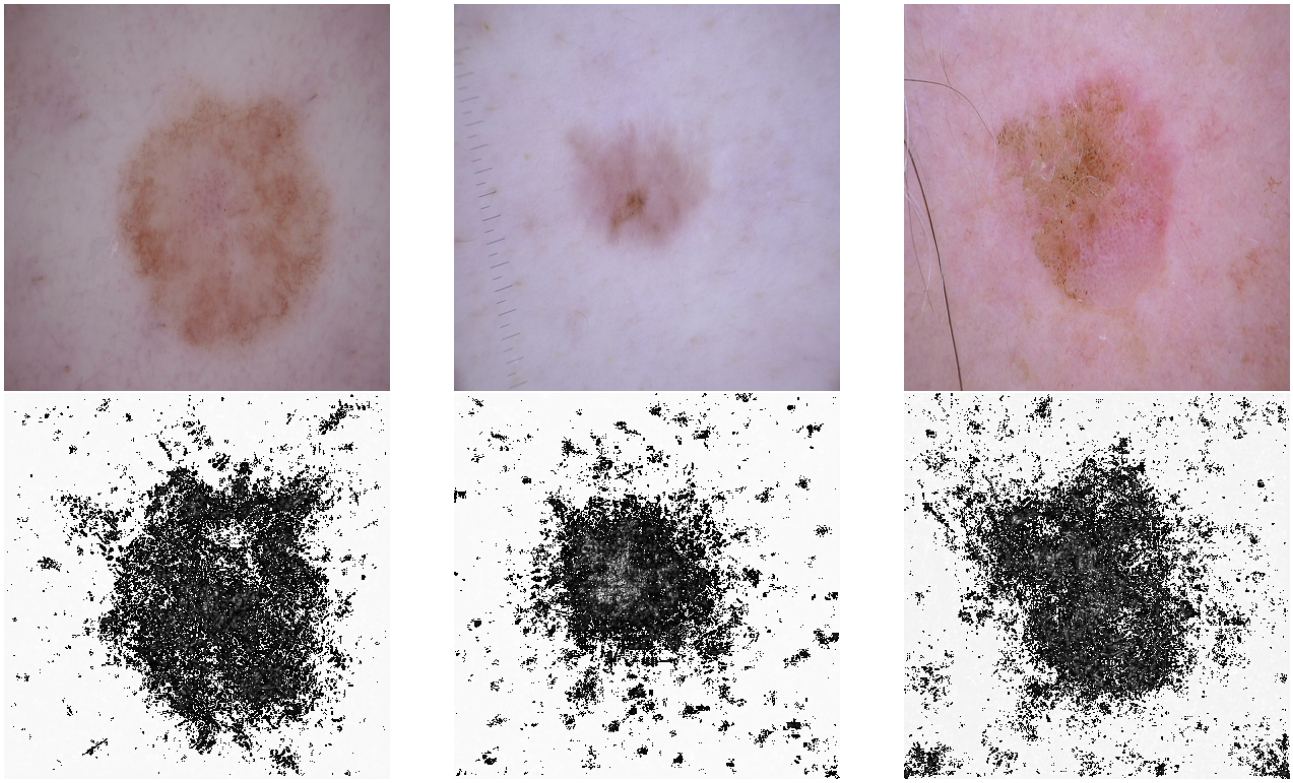


Figure 3.2 – Extracted saliency maps (bottom row) for three skin lesions of the HAM10000 dataset (top row).

3.1.3 Content-Based Image Retrieval (CBIR)

Content-based image retrieval (CBIR) aims to find the most "similar" images of a large-scale dataset against a query image. CBIR can be split into two main tasks: (i) extracting a feature representation of each input image, followed by (ii) a similarity measurement of these features. To be more specific, the task of CBIR consists of retrieving the k most "similar" images from a large-scale dataset given a query image, i.e., a new skin lesion of an unknown class. Here, the value of the cut-off k states the number of images to be retrieved. Several different CBIR techniques were developed, and it is still an active field of research. Before neural networks emerged, hand-crafted features were extracted for CBIR. The most recent techniques use two neural networks, one for extracting the features and the other for training those features regarding similarity. In this thesis, we have implemented CBIR based on deep features of a CNN trained for classification. There is a strong correlation between the retrieval and classification performances in this method due to the extraction of features of one of the last layers. Tschandl et al. have achieved a comparable multi-class accuracy of the CBIR compared to the classification outcome with this approach, where the multi-class accuracy of the CBIR was calculated using majority voting of the retrieved images [37]. A scheme of this CBIR approach is illustrated in Figure 3.3. In our model, we extract the deep features right after the last convolutional layer and before the classification layers. Due to the high dimensionality of the resulting feature space, we implemented the cosine similarity to find the query image's most "similar" images of this space. A Euclidean distance measure was tested too but produced slightly worse quantitative retrieval results. This agrees with the work of Wang et al., who showed that cosine similarity

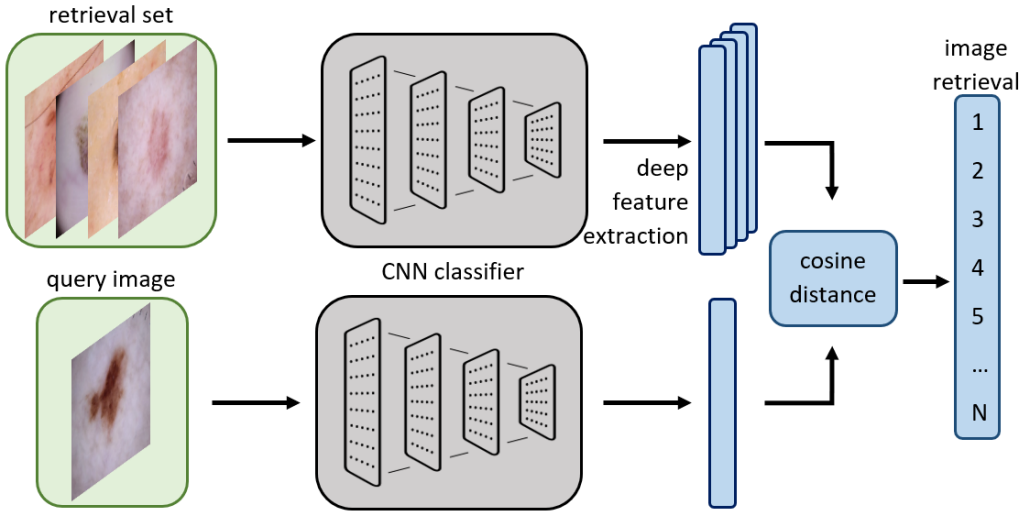


Figure 3.3 – Scheme of content-based image retrieval (CBIR) based on deep learning of a CNN trained for classification on the HAM10000 dataset.

outperforms Euclidean distance without normalization for CBIR [38]. This is most likely due to the *curse of dimensionality*, which affects the cosine similarity less, since it does not depend on the magnitude of the vectors, but only on their angles. We find the cosine similarity of the query's feature vector $\mathbf{q}$ to one of the training sets feature vectors $\mathbf{v}$ by

$$S_C(\mathbf{q};\mathbf{v}) = \frac{\mathbf{q}\cdot\mathbf{v}}{|\mathbf{q}|\,|\mathbf{v}|}. \tag{3.3}$$

3.2 Saliency-Enhanced Content-Based Image Retrieval (SE-CBIR)

A scheme of the algorithm developed in this thesis for retrieving "similar" skin lesions with respect to a query image, called *Saliency-Enhanced Content-Based Image Retrieval* (SE-CBIR), is shown in Figure 3.4. By adding saliency maps to the information flow to enhance the algorithm's focus on the seemingly important regions, we aimed to improve the retrieval of "similar" skin lesion images both quantitatively and qualitatively.
The algorithm is trained in two steps. In the first one, the 3-channel CNN classifier is trained on the HAM10000 dataset using transfer learning. In the next step, the 4-channel CNN classifier is trained by fine-tuning this first classifier on input pairs consisting of the skin lesions and their saliency maps extracted from the first classifier. In both training steps, we optimize for high sensitivity of the classification into the seven classes. More details on the training are discussed in the following subsection.
Once an optimal performance is reached, the most "similar" images to a query image can be retrieved from a large-scale dataset. Therefore, the lesions' saliency maps are extracted from the first classifier and then fed through the 4-channel classifier together with the corresponding skin lesion images. Then, the most "similar" lesions are retrieved by CBIR using the emerging deep features from the previous step. The idea behind the second classifier is to take advantage of the fact that the relevant information of skin lesion images is spatially restricted. By adding the saliency maps to the input of the second classifier, we enhance the classifier's focus on the seemingly important regions of the skin lesion images. Hence, these regions contribute

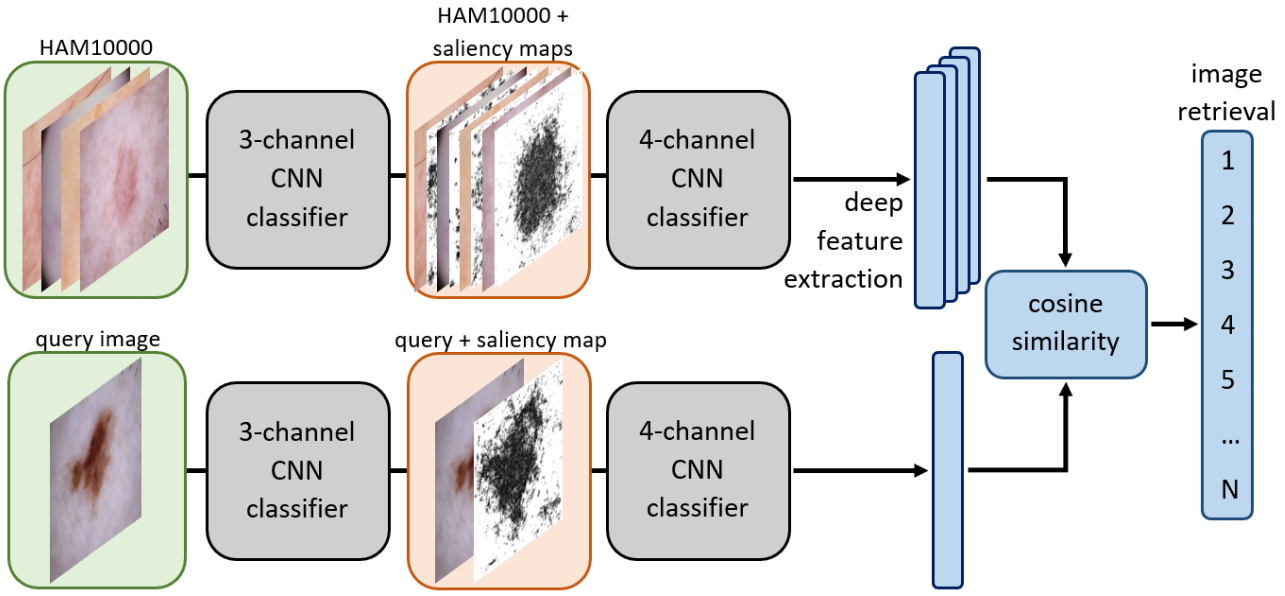


Figure 3.4 – Scheme of the Saliency-Enhanced Content-Based Image Retrieval (SE-CBIR).

intensified to retrieving "similar" images.
Our algorithm is based on the IG-CBIR algorithm of Silva et al. described in Section 1.5 [25]. We used their algorithm as a starting point, but some adaptions had to be made due to the more complex task of retrieving similar skin lesions. In more detail, we started this project with their fine-tuning approach using only the saliency maps. With this method, the information loss in the input data by switching from complex RGB skin lesion images to saliency maps might have been too large to achieve an improvement in the retrieval by this fine-tuning step. Unfortunately, switching to overlays of skin lesion images and their corresponding saliency maps did not show significant improvement, with the loss of information being the possible cause again. This led us to our 4-channel approach for fine-tuning the second classifier inspired by a similar approach for classification problems by Murabito et al. [39].

3.3 Training

The HAM10000 dataset was used for training. Unfortunately, the labels of their test set have not been released at the time of this study. Therefore, a random training-validation-test split was done, where 80% was used for training, and 10% for the validation and test set, respectively. The split was done in a stratified manner to have the same class distributions in all data sets. Throughout the whole training and evaluation, we used the same seed with a value of 71 for the random data splitting. Due to time limitations and the main focus being to include saliency maps in the workflow to improve CBIR, no cross-validation was done. The models were trained on the Piz Daint supercomputer from the Swiss National Supercomputing Center[2] in two steps: (1) training of the 3-channel CNN classifier; and (2) extracting saliency maps from the 3-channel model and fine-tuning it on image-saliency map input pairs. The two-step train-

[2]This work was supported by a grant from the Swiss National Supercomputing Centre (CSCS) under project ID sm42

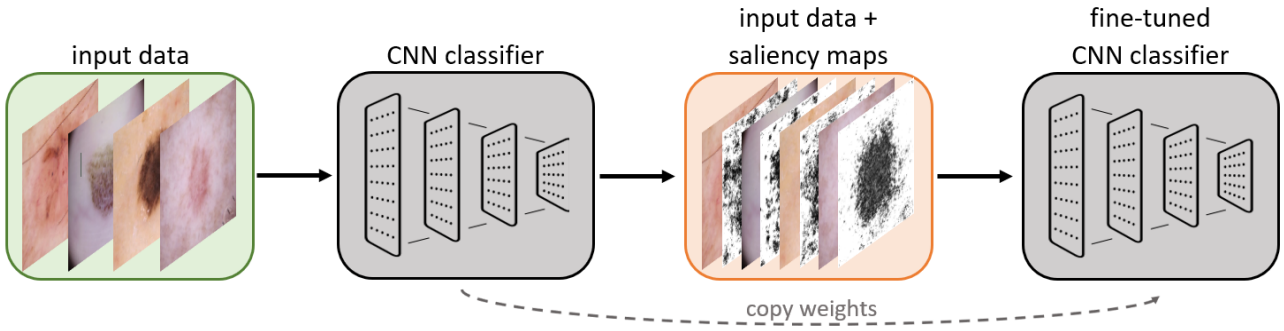


Figure 3.5 – Training scheme of the SE-CBIR algorithm.

ing procedure is shown in Figure 3.5. Before going into more detail on how the two classifiers were trained, the *adaptive learning rate* method is briefly introduced.
The main advantage of transfer learning or fine-tuning for image classification is that the pre-trained networks have already learned to recognize shapes, colors, and many more features. Several different training techniques exist to keep this information during training. All of these aim to adapt the parameters of the pre-trained layers just slightly or not at all compared to the new layers. We decided to use a so-called *adaptive learning rate* strategy in both our training steps. With this strategy, all layers are trained simultaneously but using an x-times larger learning rate for the newly added or adapted layers.

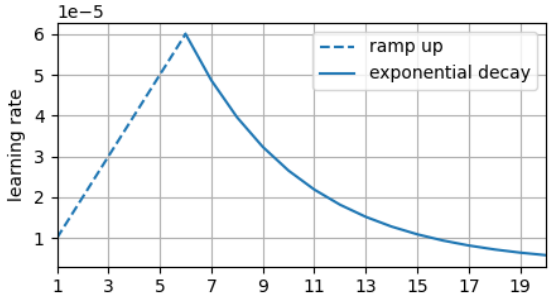


Figure 3.6 – Learning rate schedule with a ramp-up followed by an exponential decay.

Furthermore, we have used a *learning rate schedule*, illustrated in Figure 3.6. Similar to the adaptive learning rate strategy, the ramp-up in the first epochs aims to keep the learned features. After the ramp up the learning rate decays exponentially. Exponential decay is a widely used learning rate scheduling method to improve convergence.
For optimization, we have used a multi-categorical version of the focal loss [40]. This loss modifies the cross-entropy loss by an additional term, and the authors demonstrated improved performance for imbalanced datasets. With a softmax activation function in the last layer, the focal loss $\mathcal{L}_{focal}$ for each sample can be derived by

$$\mathcal{L}_{focal}(\mathbf{y}^*; \alpha, \gamma) = \sum_{i=1}^{M} -\alpha_t (1 - y^*_{t,i})^\gamma \log(y^*_{t,i}). \tag{3.4}$$

Here, variables and parameters with the subscript t are as

$$x_t = \begin{cases} x, & \text{if } i \text{ is the ground truth class} \\ 1 - x, & \text{otherwise.} \end{cases} \tag{3.5}$$

The parameters α_i and γ define the weights on this additional term, whereas y^*_i represents the softmax prediction value of an input. For $\alpha = 0.5$ and $\gamma = 0$ we end up with the cross-entropy

loss. Let us use the parameters we chose for training our models, and which the authors of the paper have also suggested: $\alpha = 0.25$ and $\gamma = 2$ for a more thorough explanation [40]. When a skin lesion image is correctly classified, $(1 - p_t)^2$ is close to zero for all classes leading to a minor addition to the loss. On the other hand, if the classification of a sample is bad, $(1 - p_t)^2$ is large for some classes leading to a significant addition to the loss function. This leads to a down-weight of well-classified cases, thus shifting the focus in training to the more difficult cases. Furthermore, the weighting factor $\alpha < 0.5$ adds more focus on the ground truth class, thus aiming for higher sensitivity during training.

Data augmentation

3-channel CNN classifier

As explained previously, a transfer learning approach using the EfficientNetB4 was used. Considering the resulting size of our CNN classifier (17 million parameters), the rather small training dataset of 8000 images, and the complex task of skin lesion classification, strong data augmentation was used to prevent our first classifier from overfitting. In this thesis, data augmentation with randomized parameters was implemented in the input pipeline while gathering a batch of images. Thus, slightly different augmentation is applied to each batch in each epoch. The implementation was done using the *albumentations* library [41]. General image augmentation techniques were used, such as

- Geometric augmentations: rotation, flips, shifts, scale on input;
- Noise: motion blur, median blur, Gaussian blur, Gaussian noise;
- Distortions: optical distortion, grid distortion, elastic transform;
- Brightness and Contrast;
- Color modifications: CLAHE (contrast limited adaptive histogram equalization), and changes in hue, saturation, and intensity values;

were applied, followed by resizing the input image, randomly cropping 10%, and applying coarse dropout. Coarse dropout is an augmentation technique that randomly removes small squares of the image to enhance the importance of the overall picture. Each augmentation technique was implemented with an intensity interval to draw a random intensity for each batch in each epoch. Figure 3.7 shows an example of a skin lesion image before (left) and after augmentation (right).

4-channel CNN classifier

We applied only minor data augmentations in this step due to the input combination of the skin lesion image and its saliency map. Besides, the network was trained for the same task already in the previous step. Therefore, random rotations and horizontal or vertical flips were applied as the only data augmentation techniques. As in the 3-channel classifier, we implemented the augmentation into the input pipeline.

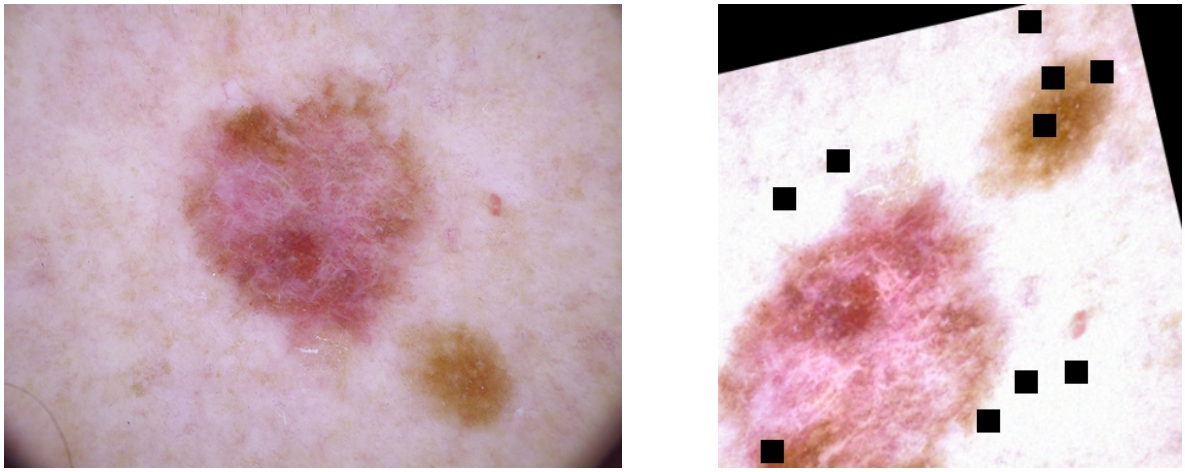


Figure 3.7 – Example skin lesion image, before (left) and after (right) applying augmentation.

Training

After 20 epochs, a good convergence for both models was reached using an adaptive learning rate strategy and a batch size of 32. Additionally, we have used the same learning rate schedule for both. This included five ramp-up epochs to a maximum value of six times the initial learning rate, followed by an exponential decay with a factor of 0.8. For the 3-channel CNN classifier, the best performance was achieved with an initial learning rate of 5e-5 combined with 200 as the adaptive learning rate factor. This led to an initial learning rate of 1e-2 for the new classification layers.
A slightly lower initial learning rate of 1e-5 with an adaptive learning rate factor of 100 led to the best performance of the 4-channel classifier. Thus, the initial learning rate for the new input layer and the first classification layer was set to 1e-3.
With these parameters, we achieved model performances as presented in Table 3.I

Model	Accuracy	Sensitivity
3-channel CNN classifier	0.79	0.59
4-channel CNN classifier	0.89	0.77

Table 3.I – Model performances: multi-class accuracy and macro-averaged sensitivity scores.

3.4 Evaluation Methods of Image Retrieval Systems

The evaluation of image retrieval algorithms can be divided into quantitative and qualitative evaluation. To begin with, the quantitative evaluation takes only the ground truths of the retrieved images with respect to a query image into account. The qualitative evaluation, on the other side, considers a combination of both the labels and the visual features of the retrieved images. For both evaluation types, several different metrics and methods have been developed. In our case, the quantitative evaluation aims to compare the retrieval of the SE-CBIR model to

retrieving images using the deep features of the 3-channel classifier with the same method. In the qualitative evaluation, on the other hand, we only aimed to examine the clinical relevance of the retrieved images by our SE-CBIR algorithm using the help of experts.

Quantitative Evaluation

In this evaluation, the performance of the retrieval is measured by taking only the labels into account.
Commonly used quantitative performance metrics of image retrieval systems are recall, precision, and f-score at k retrieved images. The formula for the leading metric, precision at k retrieved images ($P@k$), for each class i and its macro average ($AP@k$) over all M classes is shown below.

$$P@k_i = \sum_{\text{img}_i} \frac{\#\text{ of retrieved images of class i}}{\#\text{ of retrieved images}} \qquad \text{and} \qquad AP@k = \frac{1}{M}\sum_{i=1}^{M} P@k_i \tag{3.6}$$

Here the first sum goes over all query images of class i and the second over the different classes represented in the set of query images. The number of retrieved images k is often referred to as the *cut-off* value. The P@k metric was used in this thesis due to its leading role in the literature on image retrieval generally, but also in the retrieval of skin lesion images (e.g., Allegretti et al. [42] or Tschandl et al. [37]).
We used the same split for this quantitative evaluation as for training. Thus, the 1002 skin lesions of the test set, which the network has not yet seen, were utilized as query images. The training and validation sets consisting of 9013 images were used as the set to retrieve from.

Qualitative Evaluation

In many CBIR tasks, a quantitative evaluation is enough. A qualitative evaluation is necessary for tasks where classes might overlap or where the ground truths cannot be acquired with certainty. For dermoscopic datasets, unfortunately, a combination of both is present. Sometimes lesions of one type, e.g. nevi, can possess all features and characteristics of another type, e.g. melanoma, leading to an overlap of classes. Additionally, some mislabeling in a large-scale dataset, where images are collected from different sources, with different ground truth acquisition (histopathology, follow-up, expert conscious) is unavoidable. Furthermore, some lesion types, such as the so-called *unstable solar lentigo*, can develop from solar lentigo (class *bkl*) to melanoma (class *mel*) [43], leading to both class overlap and mislabeling. Therefore, a qualitative evaluation is essential to draw founded conclusions. Since the help of experts is needed for this type of evaluation, and since these are clinicians in our case, we sometimes refer to the qualitative evaluation as clinical evaluation.
The aim of the clinical evaluation is to measure the relevancy of the retrieved images with respect to the query image. Of course, relevancy is task-dependent. CBIR systems in the field of medical diagnosis aim to retrieve images that are visually close to the query and represent similar diagnosis-relevant features. For skin lesions, examples of diagnosis-relevant features are the color, shape, and structure of the lesion, or sun-damaged skin.
The normalized discounted cumulative gain (nDCG) is one possible qualitative performance metric. A set of query images (query-set) and a set of images to be retrieved from (retrieval-set) have to be selected for this metric. Experts then assign each image of the retrieval-set a

relevance value for each query image. Due to the large dataset used in this thesis and the time limitations of experienced dermatologists, which would be needed for this labeling, we decided against this metric. A small query-set and retrieval-set would lead to subjective evaluation results. Allegretti et al. did a clinical evaluation on their CBIR algorithm for skin lesion images, which seemed to be more suited for our task [42]. Their evaluation compared the accuracy of dermatologists classifying skin lesion images of two tasks. In the first task, the dermoscopic images without any additional help were provided, whereas in the second task, the five most "similar" images, according to their CBIR algorithm, labeled with their ground truths, were displayed additionally. There is one point of concern about this method compared to nDCG. The ground truths of the retrieved images are included in this type of evaluation. Therefore, there one has to be careful in drawing conclusions about the qualitative relevancy of the retrieved images. Nonetheless, for us, the benefits, i.e., the proximity to a real-world application and the higher objectivity out-weighed this negative aspect. In a meeting with medical experts from the University Hospital of Zurich, including one of their leading dermatologists, discussions about the clinical evaluation led to the following agreement on the evaluation procedure:

- Evaluation will be switched to an online interface due to the high infection rates of COVID-19 and to reach more evaluators.
- Evaluation will be executed in two tasks. In each task, 100 skin lesion images have to be diagnosed. In Task 1, no additional help is given, whereas, in Task 2, the six most "similar" images, including their labels, are retrieved by our SE-CBIR algorithm.
- Additionally, the confidence of each diagnosis by a user will be recorded using a Likert scale ranging from 1 to 5: 1: not confident at all, 2: slightly confident, 3: somewhat confident, 4: fairly confident, 5: completely confident.
- A random set of images (query-set) will be drawn for each user, where the same lesions will be used for both tasks. To reduce the bias, the images were rotated by 180° in Task 2, and the users were urged to execute Task 2 at least one day later.

The website for the clinical evaluation was created with the *Flask* python package [44]. Similar to the qualitative evaluation, we used the same split as in training our networks. Unfortunately, some images of the dataset show the same lesion but differentiate visually due to different pre-processing steps or if they have been captured at follow-up appointments. Thus, the query-set was reduced from 1002 to 649 lesions in a post-processing step. There, all duplicates from the query-set, and additionally, all queries where at least two retrieved images display the same lesion, were removed.
In Figure 3.8, an example from Task 2 of the clinical evaluation is shown. The query image is shown on the right-hand side, whereas the retrieved images are displayed next to it. For each retrieved image, a colored frame corresponding to one of the seven classes, the ground truth class name, and the retrieval rank give additional information to the user. Below these images, the user selects the class representing their diagnosis in the same color code as the frames. Additionally, they state their confidence in this diagnosis ranging from one to five before advancing to the next sample. If additional help is needed, the users can open a new window with the general information or the class definitions by pressing one of the buttons on the bottom left.

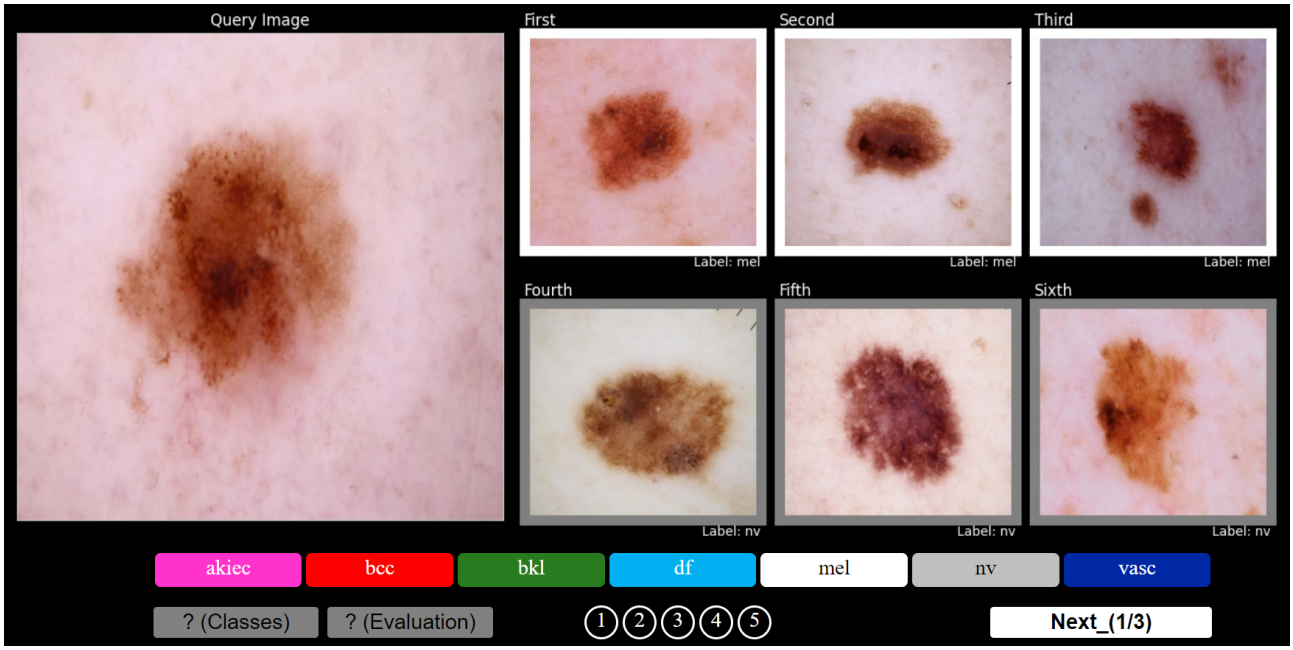


Figure 3.8 – Example from Task 2 of the clinical evaluation.

4 Evaluation Results

4.1 Quantitative Evaluation

As introduced in the previous chapter (subsection 3.4), we decided to evaluate the quantitative retrieval P@k (precision at k retrieved images) metric. Furthermore, we aimed to compare our SE-CBIR model with CBIR directly on the 3-channel classifier in order to see an improvement by adding saliency maps to the information flow.
Table 4.I shows the results of this evaluation. The $P@k$ values of each class for cut-off values $k = 1, 3, 6, 9$, and their macro-average $AP@k$ were derived by retrieving skin lesions in two ways. In the bottom row, the retrieval results from our SE-CBIR method are shown. The row above presents the quantitative retrieval results by retrieving directly from the 3-channel CNN classifier with the same CBIR method.

Model	**Cut-Off k**	**P@k**							**AP@k**
		akiec	bcc	bkl	df	mel	nv	vasc	
CBIR from 3-channel CNN classifier	1	0.69	0.75	0.80	0.55	0.67	0.95	1.00	0.77
	3	0.59	0.67	0.63	0.64	0.54	0.92	0.83	0.69
	6	0.52	0.66	0.57	0.58	0.49	0.92	0.79	0.65
	9	0.46	0.65	0.55	0.60	0.47	0.92	0.75	0.63
SE-CBIR	1	0.59	0.82	0.82	1.00	0.81	0.95	0.93	0.84
	3	0.57	0.82	0.78	0.97	0.72	0.95	0.88	0.81
	6	0.56	0.83	0.75	0.98	0.69	0.95	0.89	0.81
	9	0.56	0.83	0.75	0.96	0.69	0.95	0.90	0.81

Table 4.I – Precision at k retrieved images for each class (P@k) and their macro average (AP@k). The top values are taken from doing CBIR from the 3-channel CNN, whereas the bottom values are from our proposed SE-CBIR algorithm.

Starting with the averaged precision score (AP@k), we achieved a significant improvement (7% - 18%) for all cut-off values k with our SE-CBIR model compared to CBIR on the 3-channel classifier. Additionally, the modest decrease in AP@k with increasing the cut-off of the SE-CBIR indicates strong robustness of our model. A reason for this could be a better separation of the different classes in the feature space due to the use of saliency maps. The class-specific precision scores show stronger stability and better results as well. For all classes but *akiec* the P@k scores are significantly better, whereas for all classes but *df*, the decline with k is significantly less. The rather low scores of both CBIR models for the class *akeic* may be due to its small occurrence in the dataset (3%) combined with similar visual characteristics of solar keratoses (*akiec*) to melanomas (*mel*) and some nevi lesions (*nv*). Due to SE-CBIR's much better retrieval performance for lesions of class *df* the slight decline in P@k with k compared to a slight incline for CBIR from the 3-channel classifier is no deficiency. Furthermore, *df* is the least represented class contributing to just 1% of all skin lesion images, which adds uncertainty to these results and explains the large jump from approximately 60% P@k to almost 100% in the SE-CBIR model. The largest drop in P@k by increasing k of our SE-CBIR model is for

GitHub repository for reproducing the evaluation results.

melanomas (*mel*). The large number of nevi lesions in the HAM10000 dataset, of which many show characteristics of melanomas, could explain this drop.

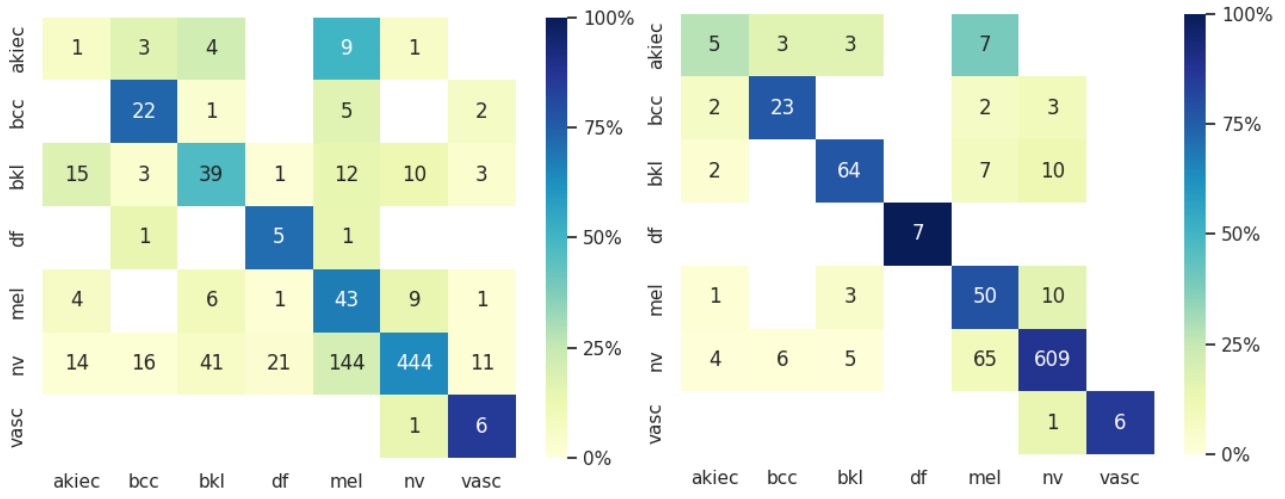
4.2 Qualitative Evaluation

The clinical or qualitative evaluation was carried out with an ad hoc online tool as described in Section 3.4. The aim of this evaluation is to see whether the retrieved images are of relevance for dermatologists.
In total, one board-certified dermatologist (PCP 1) and eight residents (PCP 2-9) from the dermatology department of the University Hospital of Zurich carried out the evaluation. The results of the two evaluation tasks (Task 1: skin lesion diagnosis without additional help, Task 2: skin lesion diagnosis with six retrieved images from SE-CBIR) and the majority vote of the first six retrieved images of the SE-CBIR are shown in Table 4.II. To avoid ties in the majority voting, weights were distributed according to the retrieval order. Since every participant evaluates on a different set of images, these accuracies differ slightly. The last row shows the average results of all participants, including the standard deviation for the diagnosis accuracy.

PCP	**SE-CBIR**	**Task 1**		**Task 2**	
#	Acc. [%]	Acc. [%]	Conf. (✓/✗)	Acc. [%]	Conf. (✓/✗)
1	89	85	3.38 (3.48/2.80)	92	4.09 (4.17/3.13)
2	90	44	2.92 (3.45/2.50)	84	4.00 (4.12/3.38)
3	89	60	2.43 (2.92/1.70)	74	2.85 (3.31/1.54)
4	92	80	2.51 (2.60/2.15)	96	3.45 (3.51/2.00)
5	89	62	2.70 (2.97/2.26)	88	4.41 (4.56/3.33)
6	86	66	4.05 (4.32/3.53)	83	4.46 (4.62/3.65)
7	92	65	3.84 (3.93/3.66)	79	4.06 (4.19/3.57)
8	93	60	3.14 (3.40/2.75)	89	4.00 (4.12/3.00)
9	88	38	3.04 (3.03/3.05)	79	3.38 (3.56/2.71)
average	89.8 ± 2.3	62.2 ± 12.7	3.11 (3.35/2.72)	84.9 ± 7.1	3.86 (4.03/2.90)

Table 4.II – Accuracy (Acc.) and average confidence (Conf.; 1: not confident at all to 5: completely confident) of 9 participants (PCP) on task 1 and task 2, consisting of classifying 100 skin lesion images with and without the support of six retrieved images by our SE-CBIR algorithm. The average confidence of correct (✓) and wrong (✗) diagnosed lesions, and the accuracy by majority voting of the SE-CBIR are shown too.

Overall, the accuracy in the diagnosis of skin lesions increased by 22%, from 62.2% in Task 1 to 84.2% in Task 2. Additionally, the average confidence of diagnoses increased from 3.11 to 3.86, however, and more importantly, the correct diagnoses contributed much more (+0.68) than wrong ones (+0.18) to this trend.
The diagnosis accuracy improved for each participant, with an improvement of significance even for participants with the highest scores in Task 1. A gain in confidence was observed for each participant, with a relatively larger confidence increase of correct to wrong diagnoses for all but participant 2.
In total, 224 wrong diagnoses of Task 1 were changed to a correct classification in Task 2, whereas just 20 diagnoses were overturned from correct to wrong. All but five of the 224 positive



(a) Accumulated results of Task 1 **(b)** Accumulated results of Task 2

Figure 4.1 – Confusion matrices showing the results of the clinical evaluation. The rows indicate the ground truths and the columns the class the lesions were assigned to by the participant. The colors indicate the proportion of the total lesion count for each class.

changes were made in agreement with the majority vote, whereas the users changed 8 of the 20 negative changes even though the SE-CBIR's majority vote agreed with their diagnosis from Task 1. This indicates that SE-CBIR influences users to change incorrect diagnoses to correct ones rather than overturning a correctly made decision by retrieving mainly images from wrong classes.

The two best performing participants of Task 1, the board-certified dermatologist (PCP 1) and participant 4, have outperformed the SE-CBIR algorithm in Task 2. This confirms the statement from above, about rather overturning wrong than correct decisions of Task 1, even for users outperforming the SE-CBIR.

Figure 4.1 shows the accumulated, class-specific results of Task 1 (Fig. 4.1(a)) on the left and Task 2 (Fig. 4.1(b)) on the right. The data is visualized in confusion matrices. These matrices can be understood as tables to describe the diagnosis performance. Each row of a confusion matrix represents the instances in an actual class (ground truths), while each column represents the instances in a diagnosed class (results from Task 1 and Task 2, respectively). To summarize, the number of cell (x, y) represents the number of diagnoses of lesions of class x (row) into class y (column). Additionally, the cells are colored according to their proportion of the total lesion count for each class. The color scale is shown on the right-hand side of each confusion matrix.

Taking a look at the diagonals first, we see an increase in correct diagnoses of each class from Task 1 to Task 2. Furthermore, the number of lesions with ground truths from other classes classified into a specific but wrong class, the so-called *false positives* or *FP*, decreased for each class. The number of non-melanomas diagnosed as melanoma (FP of melanoma) decreased from 171 in Task 1 to 81 in Task 2. The improvement for two benign classes *bkl* and *nv* are the most significant. The correctly diagnosed lesions of type *bkl* rose from less than half to 77% of correctly classified lesions, even though the total number of lesions classified as *bkl* declined from 91 to 75. The correct classifications of nevi lesions (*nv*) rose from 444 to 609, with just 80 wrongly classified nevi in Task 2. Unfortunately, the number of melanoma (*mel*) classified

as nevi did not decrease, and three basal cell carcinoma (*bcc*) were classified as nevi in Task 2, with zero of these in Task 1. Basal cell carcinomas and melanomas contribute to the most deaths of all skin cancer types, which is why misclassifications as nevi should be avoided.

5 Discussion and Conclusions

The aim of this project was to implement and test a CBIR algorithm that uses saliency maps as additional input to enhance the focus on the relevant parts of the skin lesion image. Based on the work of Silva et al. and Murabito et al. we developed a novel algorithm for the retrieval of skin lesions, Saliency-Enhanced Content-Based Image Retrieval (SE-CBIR) [25], [39]. In order to test the performance of our SE-CBIR algorithm, a quantitative and a qualitative evaluation were carried out.

The aim of the quantitative evaluation was to see an improvement in retrieval by adding saliency maps to the information flow. Therefore, the retrieval performances of our SE-CBIR model were compared to CBIR on the 3-channel CNN classifier of our model. Depending on the number of retrieved images (k), we have seen an improvement of 7% - 18% in the performance, and additionally, our model has shown strong robustness by increasing k.

The relevancy of the retrieved images of our SE-CBIR model was tested in a clinical (or qualitative) evaluation. This evaluation was carried out by a board-certified dermatologist and eight residents of the dermatology department of the University of Zurich in two steps. In the first one, skin lesions had to be diagnosed without additional help, and in a second step, the first six retrieved images from SE-CBIR including their labels were displayed as a support for their diagnosis. The participants' accuracy in diagnosing skin lesion images increased on average by 22% taking the first six retrieved images of our SE-CBIR into account. Furthermore, their overall confidence in diagnosis increased from an average of 3.11 to 3.86, where the main contribution is from more confident correct diagnoses. To our best knowledge, we were the first ones to analyze the diagnosis confidence in such an evaluation.

In a real-world scenario, this could lead to earlier detection of malignant skin lesions and fewer biopsies or follow-up appointments, resulting in a reduction of the pressure on the health care system in the realm of dermatology. Of course, this conclusion has to be treated with caution. In a clinical consultation, the participant would take more information, such as the patient's history and the context of the other lesions on their body, into account than just looking at the image of a suspicious skin lesion as we did in the evaluation. Nevertheless, these results indicate that a SE-CBIR system could be one option to close the gap between the intensive research of AI systems to augment clinical decisions for skin cancer diagnoses. One major advantage of this CBIR method is that the saliency maps could also be used as a sanity check for possible users. Since the saliency maps are part of the information flow in SE-CBIR, they are of higher significance than the postpartum extracted saliency maps of other CBIR models for a possible sanity check.

There are some limitations regarding the training and performances that have yet to be overcome or analyzed before taking the next step to a possible real-world application. First of all, we trained and tested our algorithm in a closed-loop scenario using just a single dataset for all purposes, such as training, validation, and test, but also for the retrieval and its evaluation. Training and evaluating the SE-CBIR on different dermoscopic datasets would strengthen possible conclusions. Furthermore, the use of saliency maps as part of the input data involves some risks. Many studies showed the unreliability, fragility, and inconsistency of saliency maps [45], [46], [47]. Thus, there is no certainty that our algorithm would perform as well on other datasets. Minor changes in the architecture could lead to saliency maps highlighting entirely different regions leading to possible worse results. However, the effect on the retrieval might be

minor since the initial images are used next to the saliency maps in the fine-tuning step.
Further limitations of this project regard the training and the performance of our classifiers. For training, just a single random stratified split of the dataset was used instead of several to rule out the possibility that this specific split played into our cards. Additionally, the stratified split was not ideal by just taking the class distribution into account and not the additional metadata, such as gender, age, diagnosis method, and most importantly, the lesion-id. Thus, images of the test and training set represented the same skin lesion in some cases. Cross-validation was also not applied in training. Due to simplicity, we avoided using ensemble learning, even though this training technique could have boosted our classifiers' performances. These limitations reflect the results of a submission to the ISIC 2018 classification challenge[3]. This challenge took the HAM10000 dataset as the training set and their not-yet-published test set for the evaluation. Due to the limitations stated above, especially avoiding ensemble learning, we, unfortunately, achieved significantly worse results than the winning teams.
These limitations regarding the training and performance of the classifiers mainly affect the quantitative evaluation results but should have little to no effect on the qualitative evaluation results. For the latter, the final SE-CBIR algorithm was tested on experts without any comparison between the retrieval to the 3-channel classifier. Further research has to answer the question if the quantitative evaluation scores are as promising for better-performing classifiers. On the one hand, we might see a slighter increase in quantitative evaluation scores from CBIR performed on the 3-channel model to our SE-CBIR model, where the features are extracted from the 4-channel classifier. On the other hand, the saliency maps quality increases with the classification performance, which could lead to improvements nonetheless [48].
In general, our proposed SE-CBIR algorithm is not limited to skin lesion images. Possible applications are across all image retrieval tasks, preferably where the important information is spatially restricted to benefit the most from the saliency maps. This opens the door for many future projects. The most general project would be to create a python package with our SE-CBIR such that it could be easily applied on different datasets, with different pre-trained models and saliency methods across all fields of image retrieval. Prior to this general proposal, some of the most feasible projects would be to tackle some of the limitations stated above, broaden the study to different dermoscopic datasets, and expand the clinical evaluation. Due to the online interface for the clinical evaluation, a large-scale study including many clinicians but also general practitioners could be launched. In this study, subtleties, such as adding saliency maps as sanity checks or different amounts of retrieved images, could be analyzed.

[3]ISIC 2018 Challenge: https://challenge.isic-archive.com/landing/2018/

Acknowledgements

I want to thank Dr. Javier Barranco Garcia and PD Dr. sc. nat. Stephanie Tanadini-Land for their trust, constant support and encouragement. Especially without the countless discussion with Javier, this work would not have been possible.
I would also like to thank Prof. Dr. Andreas Adelmann and his research group for supporting this project.
I am grateful for the inputs from the medical side of Prof. Dr. med, and for his help in the clinical evaluation.
Last but not least, I would like to thank my parents, my brother, my partner, and my friends for their in these last six month. Special thanks to my friends Davide, for his help in setting up the webpage for the clinical evaluation, and to Sebastián, for providing me with his LaTeXtemplate.

References

[1] Ulrike Leiter, Ulrike Keim, and Claus Garbe. "Epidemiology of skin cancer: update 2019". In: *Sunlight, Vitamin D and Skin Cancer* (2020), pp. 123–139.

[2] Philipp Tschandl et al. "Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study". In: *The lancet oncology* 20.7 (2019), pp. 938–947.

[3] Jane Scheetz et al. "A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology". In: *Scientific reports* 11.1 (2021), pp. 1–10.

[4] Pedro H Bugatti et al. "PRoSPer: perceptual similarity queries in medical CBIR systems through user profiles". In: *Computers in Biology and Medicine* 45 (2014), pp. 8–19.

[5] Marcelo Ponciano-Silva et al. "Does a CBIR system really impact decisions of physicians in a clinical environment?" In: *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*. IEEE. 2013, pp. 41–46.

[6] José Raniery Ferreira Junior and Marcelo Costa Oliveira. "Evaluating margin sharpness analysis on similar pulmonary nodule retrieval". In: *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*. IEEE. 2015, pp. 60–65.

[7] Todd W Ridky. "Nonmelanoma skin cancer". In: *Journal of the American Academy of Dermatology* 57.3 (2007), pp. 484–501.

[8] BK Armstrong and A Kricker. "How much melanoma is caused by sun exposure?" In: *Melanoma research* 3.6 (1993), pp. 395–401.

[9] Hugh M Gloster Jr and Kenneth Neal. "Skin cancer in skin of color". In: *Journal of the American Academy of Dermatology* 55.5 (2006), pp. 741–760.

[10] *Skin cancer - introduction*. Aug. 2021. URL: `https://www.cancer.net/cancer-types/skin-cancer-non-melanoma/introduction`.

[11] *Melanoma survival rates: Melanoma survival statistics*. URL: `https://www.cancer.org/cancer/melanoma-skin-cancer/detection-diagnosis-staging/survival-rates-for-melanoma-skin-cancer-by-stage.html`.

[12] Isabella Castiglioni et al. "AI applications to medical images: From machine learning to deep learning". In: *Physica Medica* 83 (2021), pp. 9–24.

[13] Alan Alexander et al. "An intelligent future for medical imaging: a market outlook on artificial intelligence for medical imaging". In: *Journal of the American College of Radiology* 17.1 (2020), pp. 165–170.

[14] Alan M Turing. "Computing machinery and intelligence". In: *Parsing the turing test*. Springer, 2009, pp. 23–65.

[15] Shane Legg, Marcus Hutter, et al. "A collection of definitions of intelligence". In: *Frontiers in Artificial Intelligence and applications* 157 (2007), p. 17.

[16] Titus Neupert et al. "Introduction to Machine Learning for the Sciences". In: *arXiv preprint arXiv:2102.04883* (2021).

[17] Mohammadhassan Izadyyazdanabadi et al. “Prospects for Theranostics in Neurosurgical Imaging: Empowering Confocal Laser Endomicroscopy Diagnostics via Deep Learning”. In: *Frontiers in Oncology* 8 (July 2018), p. 240. DOI: 10.3389/fonc.2018.00240.

[18] Weipeng Li et al. “Fusing metadata and dermoscopy images for skin disease diagnosis”. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 1996–2000.

[19] Andre GC Pacheco et al. “PAD-UFES-20: a skin lesion dataset composed of patient data and clinical images collected from smartphones”. In: *Data in brief* 32 (2020), p. 106221.

[20] Achim Hekler et al. “Effects of label noise on deep learning-based skin cancer classification”. In: *Frontiers in Medicine* 7 (2020), p. 177.

[21] Seung Seog Han et al. “Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm”. In: *Journal of Investigative Dermatology* 138.7 (2018), pp. 1529–1538.

[22] Cristian Navarrete-Dechent et al. “Automated dermatological diagnosis: hype or reality?” In: *The Journal of investigative dermatology* 138.10 (2018), p. 2277.

[23] Seung Seog Han et al. “Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm”. In: *Journal of Investigative Dermatology* 138.7 (2018), pp. 1529–1538. ISSN: 0022-202X. DOI: https://doi.org/10.1016/j.jid.2018.01.028. URL: https://www.sciencedirect.com/science/article/pii/S0022202X18301118.

[24] X Du-Harpur et al. “What is AI? Applications of artificial intelligence to dermatology”. In: *British Journal of Dermatology* 183.3 (2020), pp. 423–430.

[25] Wilson Silva et al. “Interpretability-guided content-based medical image retrieval”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 305–314.

[26] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. “The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions”. In: *Scientific data* 5.1 (2018), pp. 1–9.

[27] Ane Bonnerup Jæger et al. “Bowen disease and risk of subsequent malignant neoplasms: a population-based cohort study of 1147 patients”. In: *Archives of dermatology* 135.7 (1999), pp. 790–793.

[28] Jack M Dodson et al. “Malignant potential of actinic keratoses and the controversy over treatment: a patient-oriented perspective”. In: *Archives of dermatology* 127.7 (1991), pp. 1029–1031.

[29] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[30] Qizhe Xie et al. “Self-training with noisy student improves imagenet classification”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10687–10698.

[31] Mingxing Tan and Quoc Le. “Efficientnet: Rethinking model scaling for convolutional neural networks”. In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114.

[32] Qishen Ha, Bo Liu, and Fuxu Liu. "Identifying melanoma images using efficientnet ensemble: Winning solution to the siim-isic melanoma classification challenge". In: *arXiv preprint arXiv:2010.05351* (2020).

[33] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv preprint arXiv:1312.6034* (2013).

[34] Daniel Smilkov et al. "Smoothgrad: removing noise by adding noise". In: *arXiv preprint arXiv:1706.03825* (2017).

[35] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks". In: *International conference on machine learning*. PMLR. 2017, pp. 3319–3328.

[36] Raphael Meudec. "tf-explain: Interpretability methods for tf.keras models with Tensorflow 2.0". In: (2021). DOI: 10.5281/zenodo.5711704. URL: https://github.com/sicara/tf-explain.

[37] Philipp Tschandl et al. "Diagnostic accuracy of content-based dermatoscopic image retrieval with deep classification features". In: *British Journal of Dermatology* 181.1 (2019), pp. 155–165.

[38] Huafeng Wang et al. "Deep learning for image retrieval: What works and what doesn't". In: *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE. 2015, pp. 1576–1583.

[39] Francesca Murabito et al. "Top-down saliency detection driven by visual classification". In: *Computer Vision and Image Understanding* 172 (2018), pp. 67–76.

[40] Tsung-Yi Lin et al. "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[41] Alexander Buslaev et al. "Albumentations: Fast and Flexible Image Augmentations". In: *Information* 11.2 (2020). ISSN: 2078-2489. DOI: 10.3390/info11020125. URL: https://www.mdpi.com/2078-2489/11/2/125.

[42] Stefano Allegretti et al. "Supporting skin lesion diagnosis with content-based image retrieval". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 8053–8060.

[43] Lisa Byrom et al. "Unstable solar lentigo: A defined separate entity". In: *Australasian Journal of Dermatology* 57.3 (2016), pp. 229–234.

[44] Miguel Grinberg. *Flask web development: developing web applications with python*. " O'Reilly Media, Inc.", 2018.

[45] Pieter-Jan Kindermans et al. "The (un) reliability of saliency methods". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019, pp. 267–280.

[46] Amirata Ghorbani, Abubakar Abid, and James Zou. "Interpretation of neural networks is fragile". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 3681–3688.

[47] Richard Tomsett et al. "Sanity checks for saliency metrics". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 04. 2020, pp. 6021–6029.

[48] Taiki Oyama and Takao Yamanaka. "Influence of image classification accuracy on saliency map estimation". In: *CAAI Transactions on Intelligence Technology* 3.3 (2018), pp. 140–152.

[49] Ji Wan et al. "Deep learning for content-based image retrieval: A comprehensive study". In: *Proceedings of the 22nd ACM international conference on Multimedia*. 2014, pp. 157–166.

[50] Longlong Jing and Yingli Tian. "Self-supervised visual feature learning with deep neural networks: A survey". In: *IEEE transactions on pattern analysis and machine intelligence* 43.11 (2020), pp. 4037–4058.

[51] Ali Sharif Razavian et al. "CNN features off-the-shelf: an astounding baseline for recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2014, pp. 806–813.