UiT The Arctic University of Norway

Faculty of Science and Technology

Department of Mathematics and Statistics

Leveraging Computer Vision for Applications in Biomedicine and Geoscience

Thomas Haugland Johansen

A dissertation for the degree of Philosophiae Doctor — May 2021

Abstract

Skin cancer is one of the most common types of cancer and is usually classified as either non-melanoma or melanoma skin cancer. Melanoma skin cancer accounts for about half of all skin cancer-related deaths. The 5-year survival rate is 99% when the cancer is detected early but drops to 25% once it becomes metastatic. In other words, early detection is vital to saving lives.

Foraminifera are microscopic single-celled organisms that exist in marine environments and are classified as living a benthic or planktic lifestyle. In total, roughly 50 000 species are known to have existed, of which about 9 000 are still living today. Foraminifera are important proxies for reconstructing past ocean and climate conditions and as bio-indicators of anthropogenic pollution. Since the 1800s, the identification and counting of foraminifera have been performed manually. The process is resource-intensive.

In this dissertation, we leverage recent advances in computer vision, driven by breakthroughs in deep learning methodologies and scale-space theory, to make progress towards both early detection of melanoma skin cancer and automation of the identification and counting of microscopic foraminifera.

First, we investigate the use of hyperspectral images in skin cancer detection by performing a critical review of relevant, peer-reviewed research. Second, we present a novel scale-space methodology for detecting changes in hyperspectral images. Third, we develop a deep learning model for classifying microscopic foraminifera. Finally, we present a deep learning model for instance segmentation of microscopic foraminifera.

The work presented in this dissertation are valuable contributions in the fields of biomedicine and geoscience, more specifically, towards the challenges of early detection of melanoma skin cancer and automation of the identification, counting, and picking of microscopic foraminifera.

Acknowledgments

First and foremost, I want to thank my wonderful supervisors, Professor Fred Godtliebsen and Dr. Kajsa Møllersen. Without your guidance, experience, motivational discussions, and positivity, I never would have reached this stage of the journey. This is especially true in these past few months when I was writing this dissertation. There were many days where I felt it was hopeless that I would finish on time, but you kept motivating and pushing me towards the finish line. From the bottom of my heart, thank you!

Next, I would like to thank the two dermatologists that taught me about skin cancer and carried out the collection of hyperspectral skin lesion images; Dr. Herbert Kirchesch and Dr. Thomas Schopf. All of your assistance and insights were both crucial and formative at the start of my Ph.D. journey.

I also had the pleasure of collaborating closely with Dr. Steffen Aagaard Sørensen throughout the second half of my Ph.D. journey. Thank you for all those hours where you stared into a microscope, moving around and photographing vast amounts of foraminifera and sediment grains. Also, thank you for all of the joy-filled and insightful meetings. I look forward to future collaborations, Sir.

To all of my brilliant colleagues at the UiT Machine Learning Group; thank you for all of the great lunch discussions, silly jokes, words of encouragement, and social activities — I will never forget that bingo night at the office.

Last but certainly not least, I want to give my deepest and most heartfelt thanks to my family and friends. Without all of you, I never would have reached this day. I can never thank you enough, but I will try.

Thomas Haugland Johansen
Tromsø, May 2021

Contents

List of figures

List of abbreviations

ANN artificial neural network 21, 22

AP average precision 47

AR average recall 47

CNN convolutional neural network 22, 25, 26, 32

DNN deep neural network 22, 23

EM electromagnetic 7, 9

FCN fully convolutional network 30

HSI hyperspectral imaging 7–9

IR infrared 7

MLP multi-layer perceptron 22–24

MSE mean squared error 17

NLL negative log-likelihood 17, 24

ReLU rectified linear unit 23

RGB red, green and blue 7

RNN recurrent neural network 22

RoI region of interest 30

RPN region proposal network 30

SALSA search, appraisal, synthesis and analysis 13

UV ultraviolet 7

Chapter 1.

Introduction

There are many research fields, and associated real-world applications, that greatly benefit from advancements in computer vision driven by the deep learning revolution [1]. Two such fields of research are biomedicine and geoscience. The work presented in this dissertation is focused on skin cancer within the field of biomedicine, and microscopic foraminifera within the field of geoscience.

Skin cancer is one of the most common types of human cancer, and worldwide it accounts for around 7.9% of all reported cases [2]. Approximately 1.2% of all cancer-related deaths are attributed to skin cancer, with half of these being caused by a particular type called melanoma skin cancer. Non-metastatic melanoma skin cancer has a 99% 5-year survival rate, but this drops 25% once the cancer becomes metastatic and spreads to distant organs [3]. Therefore, early detection of melanoma skin cancer is critical to saving lives.

Foraminifera are small single-celled organisms, typically smaller than 1 mm, which are found in marine environments. During their life cycle they produce shells, referred to as tests, from various materials that readily fossilize in sediments and become part of the geological record. In total around 50 000 species have been identified, and approximately 9 000 are still in existence today [4]. By studying sediment core samples from a region, it is possible to reconstruct past ocean and climate conditions [5–7]. Foraminiferal analysis has also been shown to be valuable for detecting bio-indicators of anthropogenic pollution of marine environments [8].

1.1. Key challenges and opportunities

Deep neural networks have successfully been applied to the task of classifying skin cancer [9]. However, much work remains to be done in order to accurately detect melanoma skin cancer

at an early stage, which is crucial to saving human lives. Furthermore, it seemed to us that an upper limit had been reached with respect to what is achievable with conventional imaging methods used in dermatology. In the last decade and a half, the use of hyperspectral imaging systems has been an active field of research within biomedicine [10, 11]. One area of focus has been the detection of cancer via known bio-indicators that can be detected in specific regions of the light spectrum [12]. Based on these insights, our hypothesis is that hyperspectral imaging can be used to improve the accuracy of skin cancer detection, and at an earlier stage. Preliminary investigation uncovered that research had been published in this direction. However, when the project presented in this dissertation commenced, it was unclear what the gaps in the knowledge were, as well as what the biggest challenges in the research were. These questions combined with the promise of hyperspectral imaging in skin cancer detection constitutes the basis of our first set of opportunities;

I Critically evaluate published research conducted towards detecting melanoma skin cancer using hyperspectral imaging, and identify what remains to be done.

II Develop methods towards the early detection of skin cancer by using hyperspectral images.

Since the early 1800s, the task of identifying, counting and picking microscopic foraminifera has been done manually by geoscientists. Typically, to get statistically significant and robust representations of the fauna, a large number of specimens must be analyzed. Depending upon the complexity of the samples, and the expertise of the geoscientist, the task usually requires 2–8 hours per sample. Furthermore, a typical study often consists of 100-200 samples. In other words, the amount of time needed per study can be staggering. We state that it is necessary to develop methods towards automating the counting and picking of microscopic foraminifera; not only to reduce the time and expertise needed, but also to make studies requiring robust reconstructions of past and present faunal conditions more accessible. Important work in this direction has be done [13–16], but the challenges are not yet resolved. We define the following opportunities as first steps towards full automation;

III Implement an accurate and robust classification method for microscopic foraminifera.

IV Develop a methodology for detecting microscopic foraminifera and delineating objects with segmentation masks for fine-grained localization.

1.2. Brief overview of research

The research presented in this dissertation consists of four papers; the first two are applied to the field of biomedicine, and the final two are applied to the field of geoscience.

Paper I is a critical review of recent, peer-reviewed research on using hyperspectral imaging for skin cancer detection/classification, and addresses opportunity **I**. In the paper we first identify 86 candidate publications from the period 2003–2018, which were reduced to 20 after applying exclusion criteria based on relevance and quality of the research. The remaining 20 items of research were then critically evaluated, analyzed and synthesized. We present our findings, including critical remarks and our suggestions for future research.

In Paper II, a novel scale-space methodology for detecting very small changes in spectral signatures addresses opportunity **II**. We evaluate the method on two datasets of hyperspectral images. First we evaluate on a novel dataset of hyperspectral images of skin lesions acquired using a prototype hyperspectral camera. Because we were unable to monitor skin lesions over time (any suspected cancer is surgically removed), we induced small, artificial changes in the spectral signatures to simulate a change. To test our method without artificially induced changes, we acquired a small dataset consisting of hyperspectral images of frozen fish where images were taken at different time steps. We conclude that our scale-space methodology is able to detect changes over time.

Opportunity **III** is addressed by Paper III, where we develop a deep learning method for accurately classifying microscopic foraminifera. We first created a novel dataset of more than 2600 images of individual microscopic foraminifera and sediment grain specimens, categorized into four high-level class labels. Then we develop a deep learning classifier based on a VGG-16 [17] model with parameters that had been pretrained on the ImageNet dataset [18]. To quantify the robustness of the developed model and make it more applicable in a real-world context, we implement an uncertainty estimation algorithm.

In the final work, Paper IV, we tackle opportunity **IV** by developing an instance segmentation model using a deep learning methodology. First, we present a novel object detection dataset of microscopic foraminifera and sediment grains. The dataset consists of 104 images, where each image contains a large number of specimens that have high-quality segmentation masks. The dataset contains over 7000 objects, categorized into the same four high-level class labels used in Paper III. Second, we develop an instance segmentation model based on Mask R-CNN [19] using parameters pretrained on the COCO dataset [20]. We thoroughly analyze the model predictions, present our findings, and suggest future research directions.

1.3. Reading guide

The remainder of this dissertation is divided into four main parts. In Chapter 2, necessary background information is presented in order to give an understanding of the context in which the research is placed. We begin by giving an overview of hyperspectral imaging, before moving on to key information and statistics on skin cancer. Finally, we give some insight into the world of microscopic foraminifera.

In Chapter 3 we cover the various methodology used in the four included papers. Here we begin by describing research synthesis, more specifically, systematized literature reviews. Next, we go through several chapters covering relevant methodology from supervised learning, classification, deep learning, image segmentation, and transfer learning. Finally, we summarize uncertainty estimation within the context of deep learning, and give a brief summary of relevant scale-space techniques.

In Chapter 4–7, summaries of the four research papers are presented, which includes detailed lists of the contributions by the author. This part rounds off the dissertation with some concluding remarks in Chapter 8

The fourth and final part is Chapter 9–12, which consists of the four papers included in the dissertation.

Part I.

Background and methodology

Chapter 2.

Background

2.1. Hyperspectral imaging

Hyperspectral imaging (HSI) is an imaging technique first introduced in the field of remote sensing [21], and is based on a combination of spectroscopy and digital photography. Most digital photography systems captures data across distinct bands of the electromagnetic (EM) spectrum, where each band corresponds to the primary colors red, green and blue (RGB). Each pixel in an RGB image is assigned the recorded luminance (amount of light) values for the respective spectral bands of each of the primary colors. The goal of HSI is to assign each pixel with a very large amount of wavelength measurements, sampled uniformly across the entire bandwidth of the sensor. Depending upon the HSI system used, each pixel will generally be represented as a measurement of either absorption, reflectance, or fluorescence. Additionally, when the sampling is done using a very fine resolution, each pixel can be represented as a contiguous curve. See Figure 2.1 for the conceptual difference between RGB and hyperspectral images.

The human visual system is, on average, considered to only perceive light with wavelengths in the approximate range from 380 to 750 nanometers [22, 23]. Because the human visual system only operates in a small region of the EM spectrum, most digital photography systems are limited to the same region. This is a missed opportunity in many applications, since salient information often exists beyond the human-visible spectral range [10, 24]. HSI systems can operate beyond the visible light spectrum, in the infrared (IR) and ultraviolet (UV) spectral ranges. Additionally, due to the contiguous sampling and fine spectral resolution, techniques such as spectral unmixing and segmentation are available [25–27]. In medical applications such as cancer detection, important biological markers can reportedly be detected in the IR and UV regions [11, 12], and for cancer treatment spectral unmixing and segmentation have been used to separate cancer from healthy tissue [28].

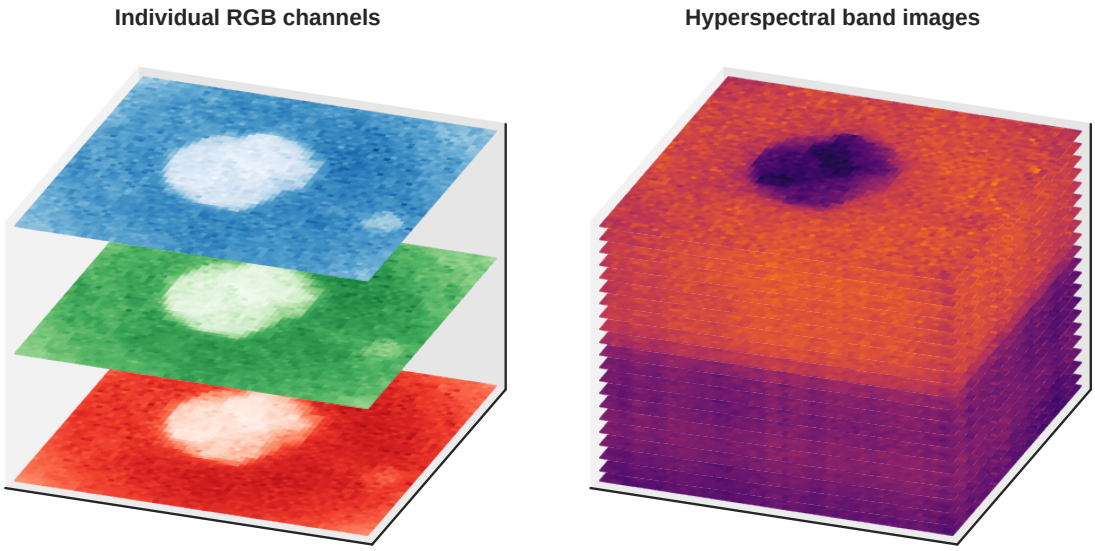


Figure 2.1. *The conceptual difference between RGB and hyperspectral images. On the left, the three channels of an RGB image are shown using false color representation of luminance values for each respective channel. On the right, a caricature of a hyperspectral cube is shown, where the original 120 spectral band images have been downsampled to 21. Each spectral band image is shown using a linear color map representation of the respective, normalized reflectance values. Both images were sourced from the same clinical, hyperspectral image of a skin lesion.*

There are four main techniques for acquiring hyperspectral images; spatial scanning [11], spectral scanning [29], spatio-spectral scanning [30], and snapshot imaging [31]. Each technique comes with its own set of advantages and disadvantages, but the end result is conceptually the same for all techniques. Hyperspectral images are often represented as a so-called hyperspectral cube, where two of the dimensions are the spatial pixel coordinates, and the third is the spectral signature. This way of organizing the data makes it convenient to operate on each set of wavelength measurements as separate images, which can be useful for many computer vision algorithms.

One aspect that is common to most HSI acquisition techniques and systems is that the raw images need pre-processing before usage [32]. Perhaps the most common pre-processing step for HSI data is converting the pixel-wise measurements to calibrated reflectance values. Reflectance represents the amount of radiant energy reflected from a surface, where 0% means all energy is absorbed and 100% means all energy is reflected. Because each HSI sensor, system, and ambient conditions vary, reflectance is almost always calibrated based on two calibration images. The first calibration image I_D, sometimes referred to as the “dark

current" image, is acquired from the HSI system when zero light hits the sensor. This image represents the minimum value for each individual pixel captured by the sensor, and corrects for inherent imperfections in the system. The second calibration image I_W is acquired by using a reference surface that reflects almost all light, and is commonly known as the "white reference" image. Importantly, the reference surface has known reflectance or absorption characteristics across a range of the EM spectrum. The white reference image represents the maximum reflectance, for the given the light emission source and the ambient conditions, captured in each pixel across the spectral operating range. Using the two calibration images, the relative reflectance image [33] can be expressed as

$$I_R = \frac{I_0 - I_D}{I_W - I_D} \in [0, 1]. \tag{2.1}$$

By calibrating images this way, we adjust for variations in light source characteristics, ambient conditions, as well as variations between different imaging systems. Making sure that images taken of the same object at different times, perhaps under slightly different ambient illumination conditions, have consistent distributions is vital for many applications such as change detection algorithms that are based on identifying statistically significant variations between observations.

2.2. Skin cancer

One of the most common types of cancer in humans is skin cancer, which accounts for about 7.9% of all cancer cases [2]. Skin cancer is typically classified as being either melanoma or non-melanoma skin cancer. In the past few decades, the reported number of melanoma cases in many countries has been increasing [3, 34]. The increasing trend does appear to have slowed in younger population groups, but is still rapidly increasing in those over 50 years old [35]. Non-melanoma skin cancer is by far the most common form of skin cancer, but is generally associated with a relatively low mortality rate. Malignant melanoma on the other hand, which is much less common, has a much higher mortality rate. Out of all new cancer cases reported in 2020, 6.2% were diagnosed as non-melanoma skin cancer, and 1.7% were diagnosed as melanoma. In the same year, 0.6% of all cancer-related deaths were caused by non-melanoma skin cancer, and 0.6% were caused by melanoma [2]. The key to effective cancer treatment is early detection, before the cancer becomes metastatic and begins spreading to other organs. Non-metastatic melanoma has been reported to have a 99% 5-year survival rate, but for metastatic melanoma, with spreading to distant organs, the 5-year survival rate drops to 25% [3]. See Figure 2.2 for a few examples of skin cancers and

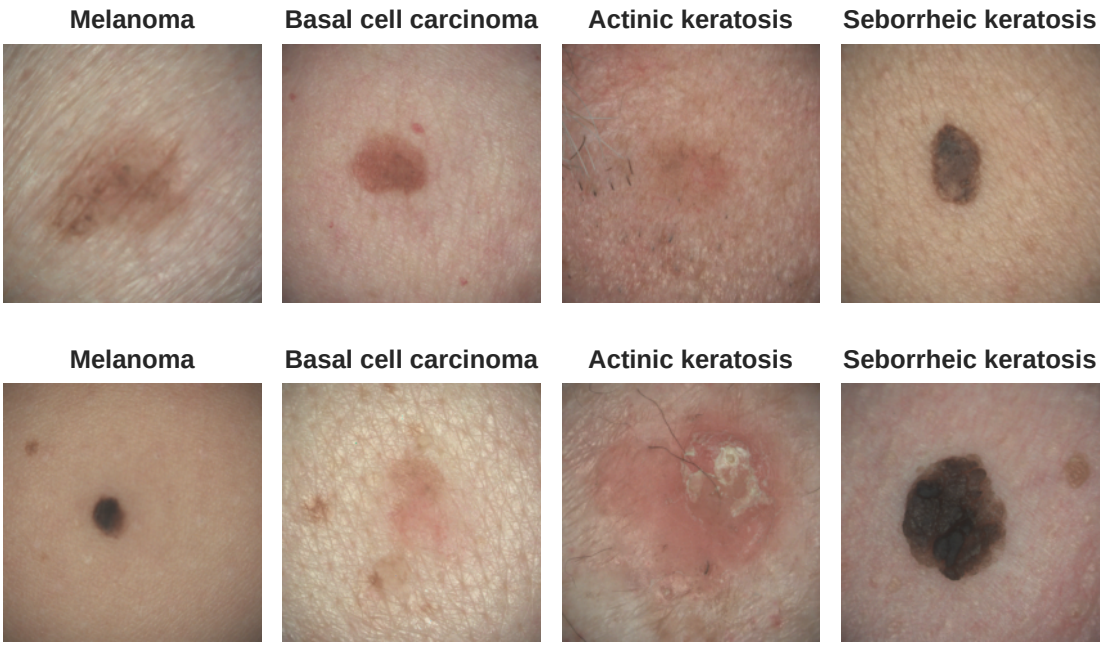


Figure 2.2. *Examples of four types of skin lesions. Both melanoma and basal cell carcinoma are classified as skin cancer, whereas actinic keratosis and seborrheic keratosis are non-cancerous skin growths. Melanomas are one of the least common forms of skin cancer, but also the most deadly by a large margin. Basal cell carcinoma is the most common type of skin cancer, however most are curable. Actinic keratosis is the most common pre-cancerous growth, and is caused by excessive ultraviolet radiation, i.e. sun exposure. Seborrheic keratosis is a harmless, non-contagious growth that generally requires no intervention.*

other skin lesions.

Physicians are frequently taught the ABCD rule of dermatoscopy [36] as a tool to assess melanocytic skin lesions for malignancy. The *A* is asymmetry, and is used to assess the symmetrical uniformity of the skin lesion. *B* is the border criteria, which captures how distinctly the lesion is delineated from the surrounding tissue. Cancerous skin tissue often exhibits non-distinct, diffuse borders in the sub-regions where there is spreading. *C* stands for color, and represents the color or mixture of colors of a skin lesion. *D* is for differential structures, and is scored based on visible structures such as pigment networks, streaks, dots, and globules. Based on the clinical assessment of each criteria, a total score is calculated, which can be used to classify the melanocytic skin lesion. The classification indicates whether the skin lesion is benign, suspicious or malignant. It is also worth noting that in recent years the ABCD rule has been extended to ABCDE, where the *E* stands for evolving. Monitoring

the evolution of a skin lesion over time has been shown to be a very important criteria for detecting skin cancer [37].

All suspicious or malignant skin lesions are surgically removed, and the excised tissue undergoes a pathology investigation. The final classification of the skin lesion is usually based on the histopathology diagnosis produced by specialized pathologists [38]. Having an accurate in-situ diagnosis is important to avoid the unnecessary surgical removal of benign skin lesions. This will help reduce the workload on pathology departments, which will in turn reduce the total wait time for determining whether further treatment or intervention is necessary. Additionally, suspected malignant melanoma must be surgically removed with a safety margin to ensure all cancer cells are removed [39]. The surgical margin can often result in large scars, and in cases where skin lesions are removed from e.g. the facial region, might lead to reduced quality of life [40–42]. Accurate and early detection of malignant melanoma is crucial to saving human lives, and reducing unnecessary strain on affected healthcare systems.

2.3. Microscopic foraminifera

Foraminifera are unicellular organisms that exist in marine environments, and most are considered to be microscopic. Foraminifera are categorized as having either a benthic or planktic lifestyle. The benthic foraminifera live on the ocean floor, and the planktic foraminifera live in the water column. Most foraminifera produce hard, external shells called tests, which are commonly constructed of calcium carbonate or agglutinated sediment particles. The shells eventually accumulate in the sediment on the ocean floor and become part of the geological record [15]. By studying sediment cores from a region, the past foraminiferal species abundance and composition can be reconstructed. Relative foraminifera abundance estimates are frequently used as proxies to study e.g. past climate conditions [43]. For the work presented in this dissertation, the objects found in typical sediment core samples are grouped into four main categories; agglutinated benthics, calcareous benthics, planktics, and sediment grains. Some examples of specimens from these four categories can be seen in Figure 2.3.

Ever since the early 1800s the identification and picking of foraminifera has largely been performed using microscopes, small needles, brushes, and other specialized tools [44]. This manual and very time-consuming process is still the standard practice for counting and estimating foraminifera abundances in most research institutions. The identification process naturally also requires special expertise, and often extensive experience, from the people

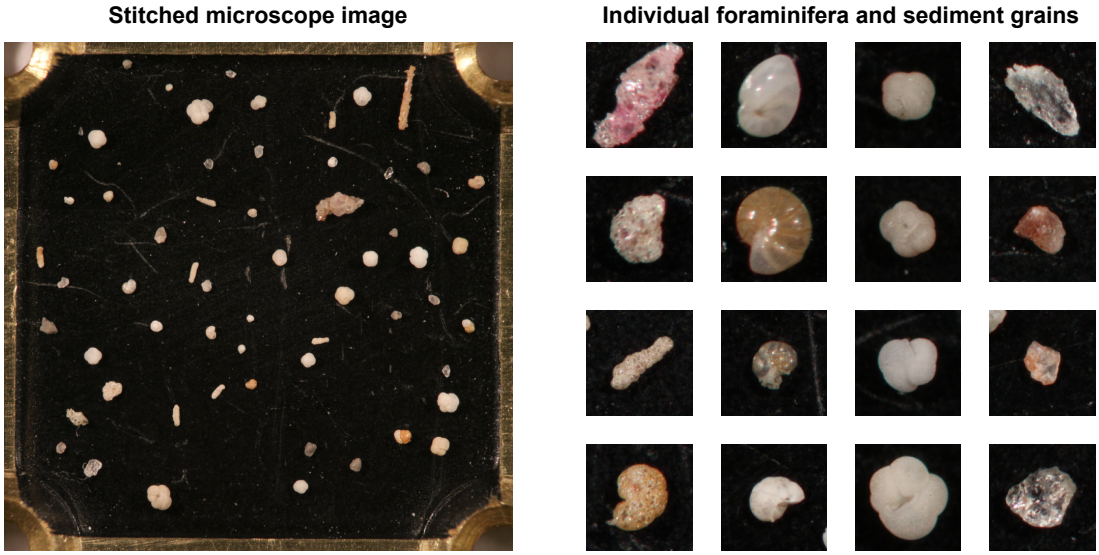


Figure 2.3. *Examples of foraminifera and sediment grain objects found in many sediment core samples. The big image on the left shows a typical super-resolution image produced by combining four different images captured from a microscope. In the image, a multitude of different types of foraminifera and sediment grains can be seen. The four columns on the right represent four high-level classes of objects that are typically seen in the big microscope images. The first column shows agglutinated benthics, the second column calcareous benthics, the third column planktic foraminifera, and the fourth column shows sediment grains.*

performing the identification. Working towards automating as much of the identification and counting process as possible is important across a multitude of research fields, such as climate reconstruction [6], and ocean acidification and pollution impact [8]. Automation will speed up research workflows and reduce the overall workload on domain experts, which will make foraminifera-based research more accessible and affordable.

Chapter 3.

Methodology

3.1. Research synthesis

The idea of the systematic review method was, most likely, first described by James Lind in 1753, a few years after his instrumental involvement in the first randomized controlled trial [45]. In the last few centuries the practice of research synthesis has become increasingly common in evidence-based practices [46]. Systematic literature reviews are perhaps most common in the field of medicine, but have also become increasingly common in other fields of research [47]. The concept of research synthesis is based on taking stock of published research within an area of interest, appraising and discarding based on some criteria, and summarizing what remains. Typically, the summary identifies what has been done, what remains to be done, the key challenges, and weaknesses in the existing body of evidence. In some cases, the research syntheses will also give new insights that were only obtainable due to a thorough, statistical meta-analysis of a large body of evidence from multiple studies [48].

For a long time, there was a lack of a consensus on how to perform research synthesis in a rigorous and reproducible manner [46, 48]. Many different approaches to research synthesis and literature reviews have been developed, each with their own subtle variations. The lack of a standard set of terminology and definitions have likely led to confusion both when reading and writing reviews. Commonly agreed upon standards for the different variations and styles of reviews do, seemingly, still not exist. With the rise of the review journals, it is common that each journal has their own prescribed definitions and methodologies. However, there have been notable publications that have systematically reviewed published reviews in order to develop common typologies. One such contribution can be found in Grant and Booth [46], which introduces the search, appraisal, synthesis and analysis (SALSA) framework. Using the SALSA framework, the authors produce a typology of 14 different types of reviews, which includes examples such as the critical review and the systematic search and review.

The critical review can be briefly summarized as consisting of extensive literature research, where each item is critically evaluated based on its quality and contribution. Based on the evaluation, items are included/excluded from further study, and what remains is presented, analyzed and synthesized. Typically, the product of a critical review is a hypothesis or "stepping stone" for new research and development, formed by taking stock of materials caught in a wide net. Another relevant review type is the systematic search and review, which can be considered as an extension of the critical review but with a much more comprehensive search process. It can, to some extent, be viewed as a hybrid between the well-known systematic review and the critical review. One of the differences from the systematic review is that the search has a much broader scope, thus giving a more complete picture. Additionally, the resulting literature is subjected to a critical review to determine its quality and value to the study. This critique process is often informal, without the necessity of a standardized method or checklist, which is contrary to the systematic review.

Regardless of the review type, the general process for conducting a review will largely remain the same and can be divided into five stages. First, the type of review and its methodology must be defined, as well as the research question and scope of the study. This stage can also include the definition of inclusion or exclusion criteria, which must then be used in subsequent stages of the review process. Second, once the scope has been defined, the search criteria need to be clearly defined, and the literature search can commence. The search criteria take into account the inclusion or exclusion criteria, if defined. Often, it might be wise to conduct some preliminary, probing search to ensure the search criteria yield satisfactory results across all relevant literature databases. Third, the published literature uncovered by the search phase needs to be assessed. The primary purpose is to assess the relevance and validity of the presented evidence, which includes study design, methodology, findings, conclusions, etc. If inclusion or exclusion criteria are defined, the assessment stage must also take these into account. Fourth, all included research must be analyzed and synthesized. The details of how this stage should be conducted will depend upon the review type and methodology, but possibly also on the defined scope. For most types of reviews, the general purpose of the analysis and synthesis is to identify gaps and limitations in the evidence. The fifth and final stage is the writing of the review, and this stage also depends on the type of review, journal requirements, author preferences, and so forth.

There are many obvious advantages of reviews, such as summarizing the current state of evidence, identifying gaps and challenges, and uncovering new insights through analysis and synthesis. However, there are also challenges related to conducting a review, e.g. avoiding introduction of bias, defining an appropriate scope, and finding all relevant published works.

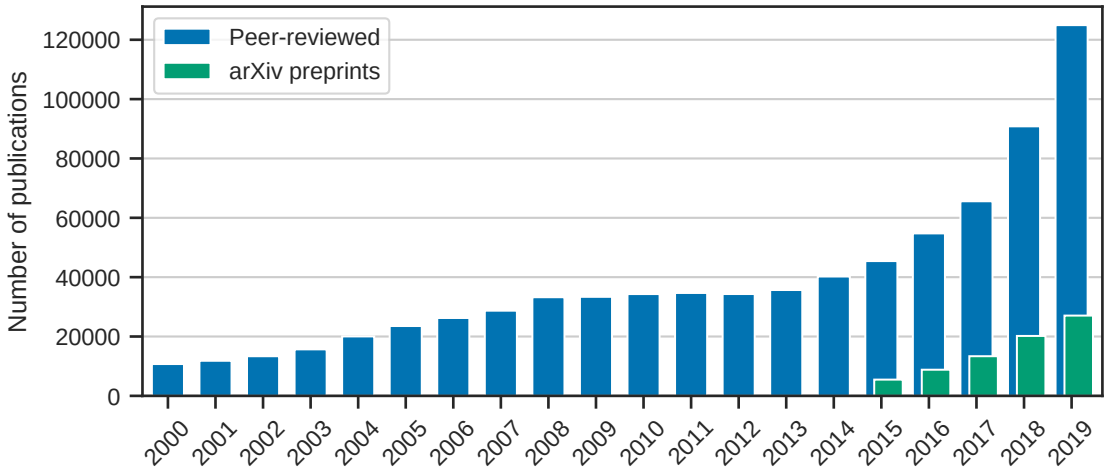


Figure 3.1. *The total number of AI-related publications from 2000–2019, which includes both peer-reviewed publications indexed by Scopus/Elsevier and arXiv preprints. The trend line for yearly peer-reviewed publications was approximately linear until around 2013/2014, which is before the deep learning "revolution." In the years 2015–2019 the growth rate for peer-reviewed publications was increasing every year, whereas the trend for arXiv preprints in the same period was linear.* Plotted using publicly available data shared via *The AI Index 2021 Annual Report* [49].

The latter can be especially challenging in fields of research that are extremely active, such as machine learning and deep learning. Also, the general consensus seems to be that reviews should only be based on work that has been peer-reviewed and published in a journal, book, or similar. But with that criteria there is always a risk that some important evidence is excluded because it has not yet passed peer-review at the time the search stage is conducted. There is also a risk that a review will be, to some extent, outdated when it is published if the dissemination process takes too much time.

The research activity within artificial intelligence, which includes machine learning and deep learning, has been increasing very rapidly in the last decade. According to numbers presented in *The AI Index 2021 Annual Report* by Zhang et al. [49], in 2000 AI-related research accounted for 0.8% of all peer-reviewed publications worldwide, while in 2013 it was 1.3%, and in 2019 it accounted for a total of 3.8%. In 2012 there were fewer than 40 000 AI publications, while in 2018 there were over 90 000 and in 2019 there were more than 125 000 publications. This means that in 2019 alone, there were on average more than 340 peer-reviewed AI-related research items published every single day. The report also presents findings from publications on arXiv, an online open-access archive for electronic preprints (often pre-peer review), from the period 2015–2020. In 2015 there were 5 478 AI-related publications, and in 2020 there

were 34 736 publications, which means a sixfold increase in 5 years. See Figure 3.1 for a year-by-year breakdown of AI-related publication statistics. Given the rapid increase in the number of publications related to machine learning and deep learning, it is becoming more challenging for researchers to stay up to date and informed. Conducting more systematic literature reviews, at varying levels of scope, target audiences, and so forth, is one obvious solution that will help alleviate the "information overload." In general, it can also be an important first publication for PhD candidates and other early-career scientists entering a new research field [50].

3.2. Supervised learning

Supervised learning describes the task of learning a function $f : X \to Y$, which transforms an input signal to its corresponding response signal. This is unlike unsupervised learning where the response signal is either not known, or for some reason, is not used. We generally assume that $X \subset \mathcal{X}$ and $Y \subset \mathcal{Y}$, where $\mathcal{X}$ is the complete input or feature space for a specific domain, and $\mathcal{Y}$ is the corresponding response or target space. Given some input $x \in X$ and output $y \in Y$, a general supervised learning task can be expressed as

$$f(x) = y, \tag{3.1}$$

where f is some parameterized function that transforms the input to the desired output. Multiple observations of input and output pairs can be organized together as a dataset

$$\mathcal{D} = \{(x_i, y_i), \dots\}, \quad \forall i. \tag{3.2}$$

Two examples of datasets are hyperspectral images where the inputs are pixel-wise spectral signatures and outputs are material property abundances, and images of microscopic foraminifera with categorical class outputs. The supervised learning objective is to somehow learn a transformation function f that generalizes to the entire training dataset. We usually also want the learned function to perform well on new, unseen datasets originating from the same data-generating process as the training dataset.

There are two main branches of supervised learning, regression and classification. In the regression setting we want to predict a continuous-valued output for a given input, whereas in classification we want a categorical prediction. Assuming we have a dataset with inputs

$\boldsymbol{x} \in \mathbb{R}^n$ and outputs y, we can approach this as a regression model;

$$y = f(\boldsymbol{x}) + \epsilon, \tag{3.3}$$

where ϵ is an error term. The error might be a byproduct of e.g. the data-generating process, manual labeling of the data, or something else that makes the dependence between $\boldsymbol{x}$ and y "noisy" to some degree. Unless the error term is deterministic or we know its distribution, learning the function f is intractable. Therefore, it is common to learn a parameterized approximation $\hat{f}$ instead,

$$\hat{y} = \hat{f}(\boldsymbol{x}; \mathcal{D}), \tag{3.4}$$

which minimizes some measure of the approximation error on the training dataset $\mathcal{D}$. In supervised learning this approximation error measure is commonly referred to as a loss function,

$$L(y, \hat{y}). \tag{3.5}$$

The choice of loss function is usually made based upon the particular task and learning algorithm, and has a big impact on the parameters of the learned function $\hat{f}$. A popular choice in regression is the mean squared error (MSE) loss, which can be defined as

$$L_{\text{MSE}}(y, \hat{y}) = \|y - \hat{y}\|^2, \tag{3.6}$$

where $y, \hat{y} \in \mathbb{R}$. For classification, the negative log-likelihood (NLL) loss is a fairly common choice, which for the binary case can be expressed as

$$L_{\text{NLL}}(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}), \tag{3.7}$$

where $y \in \{0, 1\}$ and $\hat{y} \in (0, 1)$.

Perhaps the biggest challenge in supervised learning is to learn a model that minimizes some loss on the training data $X \subset \mathcal{X}$, while also maximizing performance measures on unseen examples $X^* \subset \mathcal{X}$ and $Y^* \subset \mathcal{Y}$. One way to think about this is in terms of underfitting, overfitting, and the so-called bias-variance dilemma. An estimator or model with high bias is generally caused due the model not being capable of capturing all of the regularities in the dataset, and this can be thought of as underfitting the data. High variance is an indicator of the model overfitting to the dataset by modeling e.g. the error term ϵ, and can be caused by

the model having too much parametric freedom on the training dataset, and that the model has been optimized to the point where it has essentially "memorized" the training data. In general, we can say that a predictive model has a prediction error that can be defined as the sum of the bias error, the variance error, and the irreducible error term ϵ. Since we cannot reduce the contribution of the error term ϵ, the goal must be to somehow reduce both the bias and variance errors, but these two errors are somewhat at odds with each other. To reduce the bias error, we optimize a model using local information, e.g. fitting a line to a small neighborhood of observations instead of all observations. The variance error can be reduced by applying a smoothing function on the observations to lessen the effect of e.g. random noise. Often we rely on various regularization techniques to help reduce the bias and variance errors, in order to get models that generalize to unseen data.

A common strategy to help ensure trained supervised learning models are generalizable, is to begin by partitioning datasets into separate training and test subsets. The split ratio must be defined on a case-by-case basis, but it is good practice to always stratify the subsets to preserve the original distribution of features and labels as much as possible. The training subset is then used for learning the optimal model parameters, while the test subset is only used to measure the task performance of the trained model, i.e. how well the model generalizes to unseen examples. A strategy like this works remarkably well, but does require that the original dataset contains enough examples to yield sufficiently sized subsets. If the training subset is too small, learning optimal model parameters, with respect to task performance measures on the test subset, will be challenging. If the test subset is too small, the tasks performance measures are less likely to be good indicators of the generalizability of the trained model, with respect to other unseen examples. Both subsets need to be large enough to capture the characteristics of the distribution of the feature space $\mathcal{X}$ and label space $\mathcal{Y}$.

Another common practice is to introduce a third partition of the dataset, often referred to as the validation set. A validation set can be produced by splitting the training set split using some desired partition ratio, and this split should in general also be stratified. During training, the validation set can be used to evaluate the model performance in order to monitor when the model begins overfitting to the training set. The evaluation with the validation set can be performed at the end of every training epoch or some other desired epoch interval. By comparing the training loss with the task performance on the validation set, it is possible to identify when the model has likely begun to overfit. A popular regularization technique often referred to as early stopping is implemented by monitoring the divergence between training loss and evaluation metrics; after some predefined number of epochs of increasing divergence between the two, the training procedure is stopped. Sometimes the model parameters are

also reverted to their state before the overfitting phase was detected.

If the validation splits are randomized at the start of every training run, k-fold cross validation [51] can be implemented by repeating the training k times. This produces k different models and validation sets, which makes it possible to estimate model predictive mean and uncertainty, as well as the bias and variance in some cases. Additionally, the k different models can be considered an ensemble of models, such that ensemble predictions can be produced with e.g. majority-vote or some other ensemble approach.

The choice of model is important, and there is an almost endless pool from which we can pick and choose, but this can in itself be a source of overfitting. In other words, searching for a model that performs well on the training data is a form of overfitting, and this can often be the case with large, complex models with lots of capacity. Without prior knowledge, the first choice should therefore usually be a simple model, rather than a complex one, because it is less likely to overfit since it has less capacity to do so. In a regression setting it is not uncommon to begin experimentation with a simple linear regression model, even when the assumption is a non-linear relationship between variables;

$$\hat{y} = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + \epsilon, \tag{3.8}$$

where ϵ is an error term, sometimes referred to as a residual.

3.3. Classification

Classification is a supervised learning task that can be described as learning a mapping from input examples to their respective, categorical labels. For a m-class classification task, the categorical output is typically either given as a scalar class label, meaning $Y \in \{0, 1, 2, \ldots, m\}$, or as a set of m per-class probabilities.

The classification mapping function $f : X \to Y$ is often referred to as a classifier or classification model. The simplest type of classification algorithm is a binary classifier, which produces an "either/or" class prediction, e.g. $f : X \to \{0, 1\}$. Binary classification can be used for problems such as classifying hyperspectral images of pigmented skin lesions as melanoma or non-melanoma cases. Any linear regression model (3.8) can be turned into a simple binary classifier by thresholding output predictions. We can define a threshold function that assigns

class labels based on predictions being above or below a defined threshold τ;

$$f(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{w}^\mathsf{T}\boldsymbol{x} + b > \tau \\ 0 & \text{otherwise} \end{cases}, \tag{3.9}$$

where $\boldsymbol{w}$ are the parameters from the linear regression model, and b is a threshold term often referred to as a bias. Two concerns with this approach: (i) the regression model has not been trained to find the best linear discrimination, but rather to find the best linear correlation between input and output variables, and (ii) there might not be an obvious or intuitive threshold value for achieving good classification accuracy on unseen examples. An alternative, and more well-suited, approach to binary classification through regression is to use logistic regression. Logistic regression is a simple, yet effective and popular binary classifier, which is frequently seen in many fields of research, e.g. biomedicine [52–54]. Unlike a linear regression model, a logistic regression model yields predictions that can be interpreted as probabilities through the use of a logistic function $\sigma : \mathbb{R} \to (0, 1)$,

$$\sigma(x) = \frac{1}{1 + \exp(-x)}. \tag{3.10}$$

This type of function is sometimes referred to as a sigmoid function due to the characteristic "S" shape of its curve. See Figure 3.2 for two examples of sigmoid functions; the logistic function and the hyperbolic tangent. By construction, the logistic function assigns any negative input to the range $(0, 0.5)$, positive input to $(0.5, 1)$, and a zero input to 0.5. To use logistic regression for classification, categorical (binary) predictions can be made by thresholding the predicted probability at some defined value, e.g. 0.5;

$$\hat{y} = f(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \sigma\left(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} + b\right) > 0.5 \\ 0 & \text{otherwise} \end{cases}. \tag{3.11}$$

Like for many general regression models, there is no closed-form expression for the optimal parameters $\boldsymbol{w}$, and therefore they must be found using an optimization algorithm; maximum likelihood estimation is a popular choice for logistic regression [55].

The perceptron is an algorithm used for training a binary classifier, and was introduced by Frank Rosenblatt in 1958 [56]. In its simplest form, the perceptron algorithm learns the

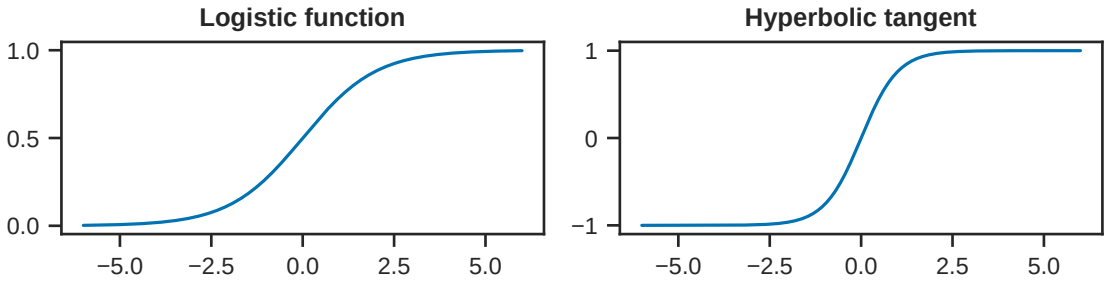


Figure 3.2. *Examples of two sigmoid functions.*

parameters of a variation of the threshold function (3.9),

$$\hat{y} = \operatorname{sign}\left(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + b\right) = \begin{cases} 1 & \text{if } \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + b \geq 0 \\ -1 & \text{otherwise} \end{cases}, \tag{3.12}$$

where the true labels are now defined as $y \in \{-1, 1\}$. When generalized to include an activation function, such as the logistic function (3.10), it is sometimes referred to as an artificial neuron. Interestingly, with the logistic activation, the perceptron is essentially an implementation of logistic regression [57]. The generalized form of the perceptron is one of the basic building blocks of artificial neural networks (ANNs). When using a linear activation function, and the observations in the dataset are linearly separable, the perceptron algorithm is guaranteed to converge [57]. Note that this strong convergence guarantee does not hold when the activation function is nonlinear. There are many ways to implement the learning procedure for a perceptron; one approach is the so-called delta rule, which is a gradient-based optimization technique [58]. See Figure 3.3 for a simple illustration of a delta rule update of a perceptron being trained in an online scheme. Here, online means that the parameters are being updated as new examples are being fed to the algorithm, one example at a time.

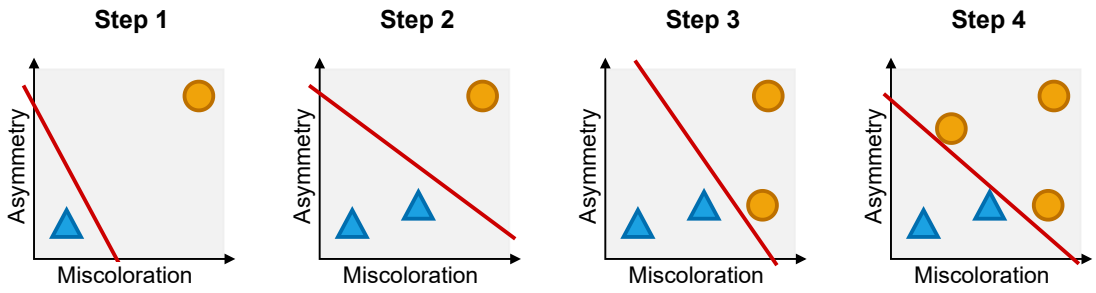


Figure 3.3. *Illustration of the perceptron algorithm. The parameters are updated using an online scheme, which means that parameters are updated as new examples are presented to the algorithm.*

3.4. Deep learning

Deep learning is a sub-category of machine learning and consists of supervised, unsupervised, and semi-supervised algorithms that are typically trained on very large datasets. The most well-known and widely used deep learning algorithms are the so-called ANN. More specifically, the deep neural network (DNN), recurrent neural network (RNN), and convolutional neural network (CNN). The training of deep learning models is generally performed using some type of gradient-based optimization, and often involves careful hyperparameter tuning, regularization methods, etc. Deep learning algorithms have been around for a while, but historically they were difficult to apply to practical problems, in part due to a lack of sufficiently large, high-quality datasets, computational power, as well as several important algorithmic breakthroughs [59].

The first type of ANN was the feedforward neural network, which is based on the idea of combining multiple perceptrons together as a directed, computational graph, where the input to the network is only passed "forward" through the graph, and producing outputs at the leaf nodes of the graph [55]. Frequently, this type of network is referred to as a multi-layer perceptron (MLP). The underlying concept of deep neural networks is function composition, represented by perceptrons/neurons, where each function in the chain is referred to as a network layer. Each layer in the network represents an affine transformation, typically combined with a nonlinear activation function such as a sigmoid. The nonlinear activation gives the network increased capacity, and allows it to learn complex relationships between features that is not possible with a linear model [58]. A neural network with three layers can

be expressed as

$$f(\boldsymbol{x}) = (f_1 \circ f_2 \circ f_3)(\boldsymbol{x}) = f_3(f_2(f_1(\boldsymbol{x}))). \tag{3.13}$$

It is common to refer to the function f_1 as the first hidden layer, f_2 as the second hidden layer, and f_3 as the output layer. The term hidden layer comes from the realization that these layers learn latent feature representations of the input, not found in the training data. Instead, during training the neural network must learn the parameters of these hidden layers such that the predicted outputs match the desired outputs as closely as possible. This approach of composing together many relatively simple (nonlinear) functions, gives deep neural networks incredible capacity towards approximating target functions. In fact, the *universal approximation theorem* states that an MLP with a single hidden layer can approximate any continuous function (with compact support) up to an arbitrary level of accuracy, when the number of units in the hidden layer goes to infinity [60–63].

The choice of activation function in each layer of a network is crucial because it defines a new feature representation of the input, and thus controls how much information each unit can express. The function naturally also controls the domain of the output, which is an important consideration when designing each layer of a DNN. For example, if the activation function is the logistic function (3.10), the input will be log-transformed and the output is constrained to $(0, 1)$. In the past, the sigmoid family of functions, e.g. logistic function and the hyperbolic tangent (Figure 3.2), were the de facto standard in DNNs. More recently however, the rectified linear unit (ReLU) [64, 65],

$$g_{\text{ReLU}}(z) = \max\{0, z\}, \tag{3.14}$$

has become the recommended default activation function in most applications [55]. One of the primary reasons for this shift is that the gradients of sigmoid activations tend to saturate for very deep networks, but this is not the case for the ReLU function. Vanishing gradients causes challenges for most gradient-based optimization methods with respect to updating model parameters.

Activation functions naturally also have an important role in the output layers of networks, since this is where the predicted outputs y are formed. There is an important interaction between the choice of loss function and the activation function of output units, so the choice must be well informed. In a regression setting it is common to use a simple linear activation function in the output layer, often with a mean-squared error loss. In the case of binary classification, a sigmoid function can be used, and for classification with m classes the softmax

function is a very common choice. The softmax activation for the i-th class can be written as

$$g_{\text{softmax}}(\boldsymbol{z})_i = \frac{\exp\{z\}_i}{\sum_m \exp\{z_m\}}, \tag{3.15}$$

where the inputs $\boldsymbol{z}$ are unnormalized log probabilities predicted by a linear activation layer. The function is evaluated element-wise on $\boldsymbol{z}$ to get the softmax activation for each of the m classes, and the outputs are often called softmax probabilities since $z_i \in [0, 1]$ and $\sum_m z_m = 1$.

Most feedforward neural networks are trained using the backpropagation [66] algorithm, which can be summarized as a two-step procedure: (i) training examples are presented to the network and a prediction is produced, which is then used to calculate the loss with respect to the true target, and (ii) the parameters of each layer are updated in reverse using gradient descent, starting from the output layer and ending at the input layer, such that the loss function is minimized. In other words, the errors made at each layer, with respect to the loss function, are reduced. The key insight of the backpropagation algorithm is that by moving backwards from the loss, is that the chain rule of derivation can be exploited, which greatly reduces the computational complexity. Now it should be more clear, as alluded to before, why saturating gradients are a problem in a gradient-based optimization scheme; if the loss function is flat, the gradients become small, which undermines the ability of the network to update weight parameters. In practice, most neural networks are trained using maximum likelihood, which results in a NLL loss function. Two key benefits are that it helps reduce the chance of saturating gradients, and it simplifies the task of constructing well-behaved loss functions for each model [55].

The gradient descent procedure of the backpropagation algorithm is generally computed using stochastic gradient descent. Stochastic gradient descent avoids the need to calculate the true gradient of a loss function, by using (small) randomly sampled subsets of the training data. By using smaller batches of training data, an approximate gradient step is taken instead, which greatly reduces the computational complexity and makes it possible to train on very large datasets. There are many popular variants of the original stochastic gradient descent algorithm, such as Adam [67] and Nesterov momentum [68].

One challenge with MLP networks that contain a large number of neurons spread across many layers is the number of parameters that need to be learned. The large number of parameters is a side-effect of the fully-connected nature of these networks, since each connection has an associated weight parameter that must be optimized. Consider for a moment an example in which we want to train an MLP image classifier on the ImageNet dataset, where we have rescaled and cropped all images to 256×256 pixels. And for this task we want to predict

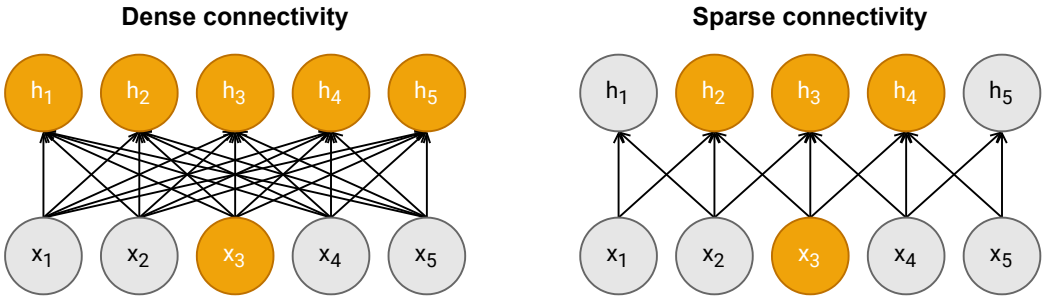


Figure 3.4. *Examples of both dense and sparse neuron connectivity.*

class probabilities for each of the 1000 class labels in the dataset, so our output layer will have 1000 units. With these input and output dimensions, and with only a single hidden layer of 4096 units, our simple model would have more than 272 million parameters. Even with this many parameters, it is unlikely this simple model would be able to learn general feature representations for the more than 1.2 million images in the ImageNet dataset [18]. Conversely, many of the popular deep CNN models that have achieved state-of-the-art results on ImageNet have vastly fewer parameters [69–71]. This is in part due to clever exploitation of the structure and statistical properties of images, and concepts such as weight sharing, feature pooling, and so forth.

One of the key differences between a fully-connected neural network and a CNN is that inner-products between weights and inputs are replaced by convolutions. Instead of assigning a single weight parameter to every pixel in the input image, a shared weight matrix is applied to the entire image using the convolution operation. In other words, we go from dense connectivity (every output is connected to every input) to sparse connectivity, which reduces memory requirements and computational complexity. See Figure 3.4 for a simple example illustrating the difference between dense and sparse connectivity. The shared weight matrix is often referred to as a filter or kernel, and a set of filters are sometimes referred to as a filter bank or feature map. Depending upon how the coefficients of the filter are defined, and how the filter is applied to the image, different (local) image features can be extracted. Examples of filters can include different types of edge detectors, texture detectors such as Gabor filters [72, 73], and so forth. Figure 3.5 shows an example of a very simple vertical edge detection filter applied to an input image using convolution. A very important consequence of learning filters with parameters that are shared across an entire input, instead of traditional fully-connected layers, is that it is a form of regularization. By imposing that a relatively small number of "tied" filter coefficients must learn to extract salient features, the complexity of the mapping function in each layer is reduced. Importantly, reducing the complexity of

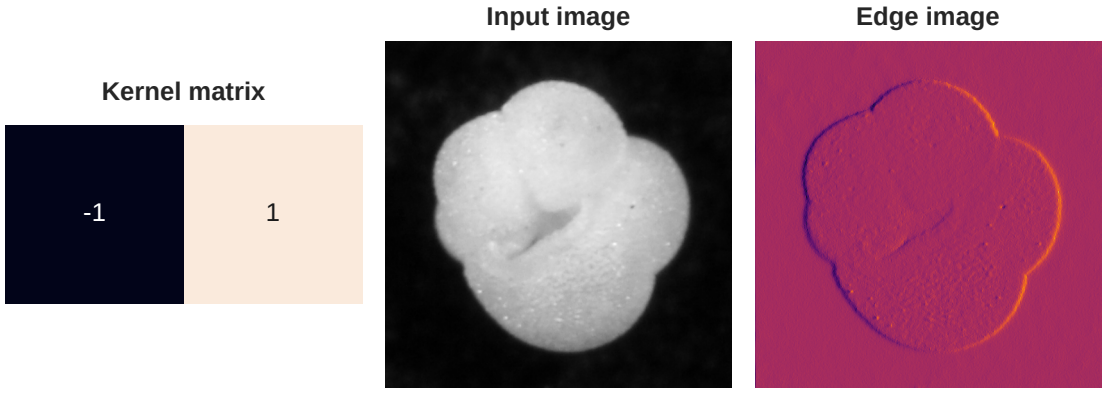


Figure 3.5. *Example of an edge detection filter applied to an image using convolution.*

the learned function is akin to preferring simpler models over complex models, which can reduce the chance of overfitting.

In traditional digital image processing applications, filters are hand-designed to extract desirable features from an image, but in a CNN the filter coefficients are learned. Each layer of a CNN learns its own feature map, and the features extracted from each layer used as input to the next, which gives each layer an increasingly high-level summary of the image. An important aspect of the increasingly high-level features is that the so-called receptive field of each feature map also increases. By receptive field we refer to the size and location of regions of the input features the extracted features in each layer attends to. The very first layer typically only attends to small local regions, just large enough to detect e.g. oriented edges, textures, and similar. Towards the end of the network, the extracted features will attend to larger regions of the image, and be sufficiently high-level to detect the presence of objects [74]. The size of the receptive fields depends upon several factors such as the size of the filter, the stride of the convolution, and pooling operations. See Figure 3.6 for a simple illustration of the receptive field concept.

The filters learned in a CNN model are translation invariant, which means that any shift in the input results in an equivalent shift in the output. This property has both advantages and disadvantages, but for computer vision tasks we might not always care about the exact location of a specific feature. One way to reduce the effect of the translation invariance is to perform what is typically referred to as feature pooling. By pooling feature activations we are essentially producing a summary-level feature of a neighborhood of feature activations. As an example, if we only care about the “strongest” feature activations per neighborhood in an input we could perform max pooling. There are many other types of pooling operations used in practice, but max pooling is the one that is perhaps most commonly used. See Figure 3.7

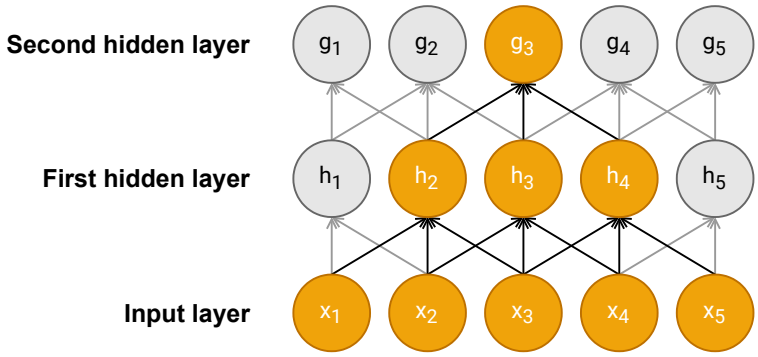


Figure 3.6. *Illustration of the concept of the receptive field in a CNN model. The input to each hidden layer is calculated using a one-dimensional convolution using a filter width of 3. This means that each unit in the hidden layers is connected to three units in each preceding layer, respectively. Ultimately, this means that each unit in the second hidden layer has a receptive field wide enough for every feature in the input layer. If the input layer was wider, we would have to e.g. increase the depth of the network, the width of the filter, or use pooling to increase the receptive field accordingly.*

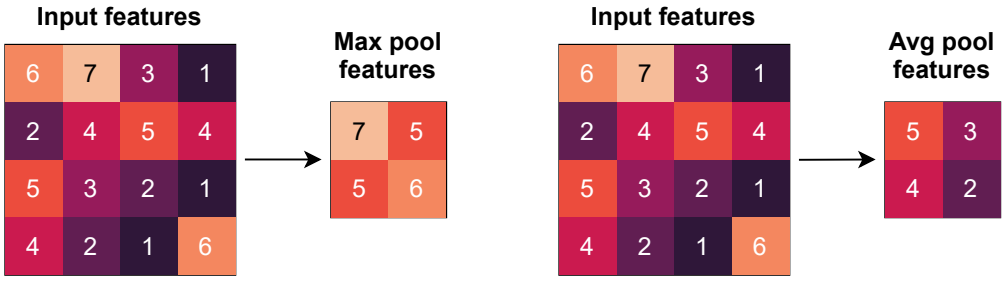


Figure 3.7. *Illustration of two-dimensional pooling operations on input features. Both pooling operations are performed using a 2×2 neighborhood with a stride of 2, which means there are no overlapping between neighborhoods. From each of the four neighborhoods, either the maximum value or the average value of the neighborhood is used to form the pooled output features. Max pooling captures the most "important" feature activation in an input region, and is the most commonly used pooling operation in computer vision.*

for an illustration of max pooling and average pooling. The size of the neighborhood is a hyperparameter that is chosen when the model architecture is designed; for computer vision a 2×2 pooling neighborhood is very common and reduces the size of the affected output feature dimensions by half. As mentioned before, the size of the pooling neighborhood also affects the size of the receptive field.

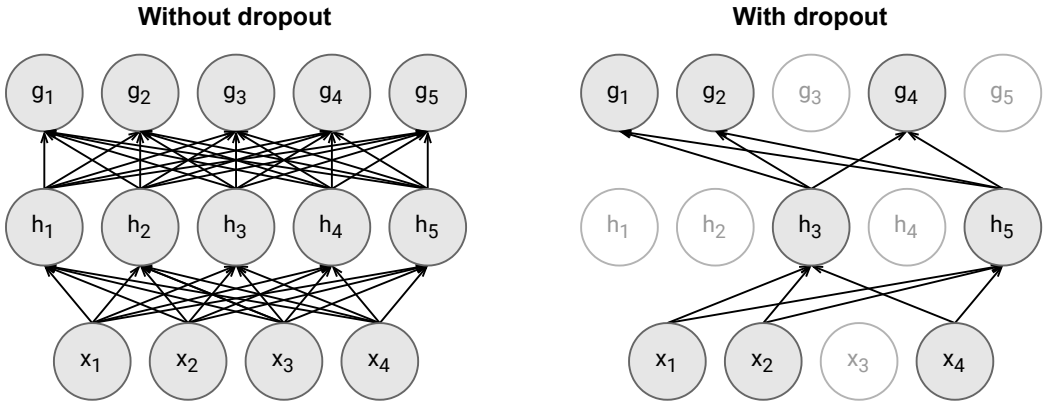


Figure 3.8. *Illustration of dropout regularization. The two networks are the same, but on the left no dropout regularization is employed, which is equivalent to a drop probability of 0. On the right, dropout is applied on the input layer with drop probability of 0.25, and on the final two layers with a probability of 0.5.*

When training deep neural networks, their tremendous capacity often requires the use of regularization techniques to prevent overfitting. Regularization becomes particularly important when datasets are considered small with respect to the complexity of the model; with enough capacity, the weight parameters can end up essentially memorizing the training examples through co-adaptation of feature detectors. One of the most frequently used regularization methods for combating co-adaptation is known as dropout [75]. The idea of dropout is to randomly "drop" incoming features to a neuron, with a defined probability, and thus forcing each neuron to learn meaningful features. Informally, with a dropout probability of 0.5, half of the units in a neuron will be switched off, and which units are affected is randomized (typically for each training batch). In which layers to introduce dropout, and what dropout probability to assign to each layer, must be chosen when designing the model. An illustration of the dropout regularization technique can be seen in Figure 3.8. Another example of regularization for reducing overfitting is random data augmentation. In computer vision applications this often includes augmentations such as additive noise, horizontal or vertical flipping, rotation, resize and crop, and adjustments to contrast and brightness. To some extent, random data augmentations expands the effective size of the training dataset.

3.5. Image segmentation

Image segmentation is the task of dividing an image into multiple segments based on some criteria or measure, usually at the pixel level. The goal of image segmentation is generally to

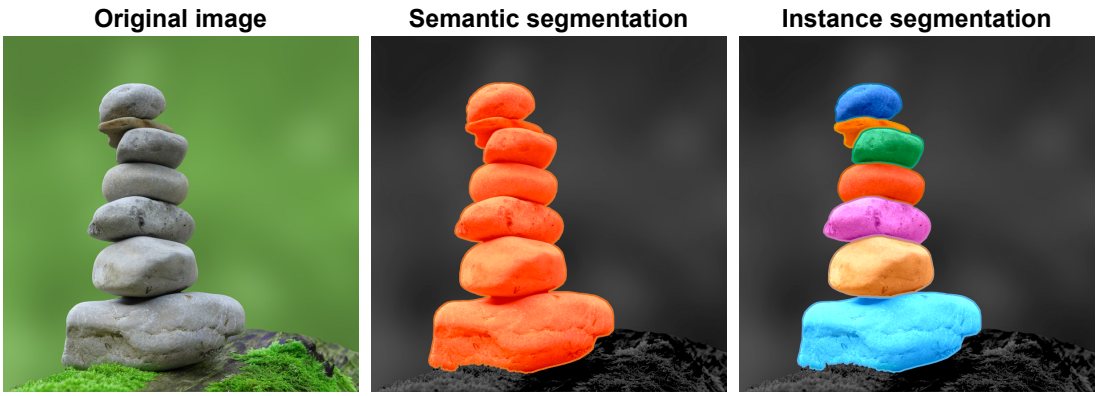


Figure 3.9. *The conceptual difference between semantic segmentation and instance segmentation. In this simple example there are two semantic classes of objects; "stone" and "background." Both segmentation maps have found the same type of objects, but the instance segmentation has also detected the distinct instances of the "stone" class.*

produce a higher-level representation of an input image, which can be useful for downstream tasks such as cancer detection [76], self-driving cars [77] and visual question answering [78].

Image segmentation methods are generally divided into two categories; semantic segmentation and instance segmentation. Semantic segmentation can be summarized as the task of assigning class labels to each individual pixel in an image. Classifying pixels in an image is useful in many applications, e.g. to delineate regions of skin as cancerous or non-cancerous tissue [79]. Instance segmentation gives each pixel in an image an object assignment, typically based on some similarity or affinity measure. This means that instead of trying to figure out what category the pixel belongs to, the task is rather to figure out which of the objects in the image a pixel belongs to. See Figure 3.9 for a visual depiction of the conceptual differences between the two types of image segmentation.

Object detection, which can be described as the task of detecting individual instances of semantically segmented objects, can conceptually be solved by naively combining both semantic segmentation and instance segmentation. We might envision that this approach could be solved as a two-step process, where we first semantically segment the image by assigning class labels to every pixel, and then determining the distinct object assignments within the semantically segmented pixels. This naive approach assumes that objects have a single semantic class, but if that is not the case the order of segmentation could be changed. However, in deep learning, object detection and segmentation are solved using more sophisticated approaches, and both tasks can be solved in parallel in a single forward-pass of a neural network.

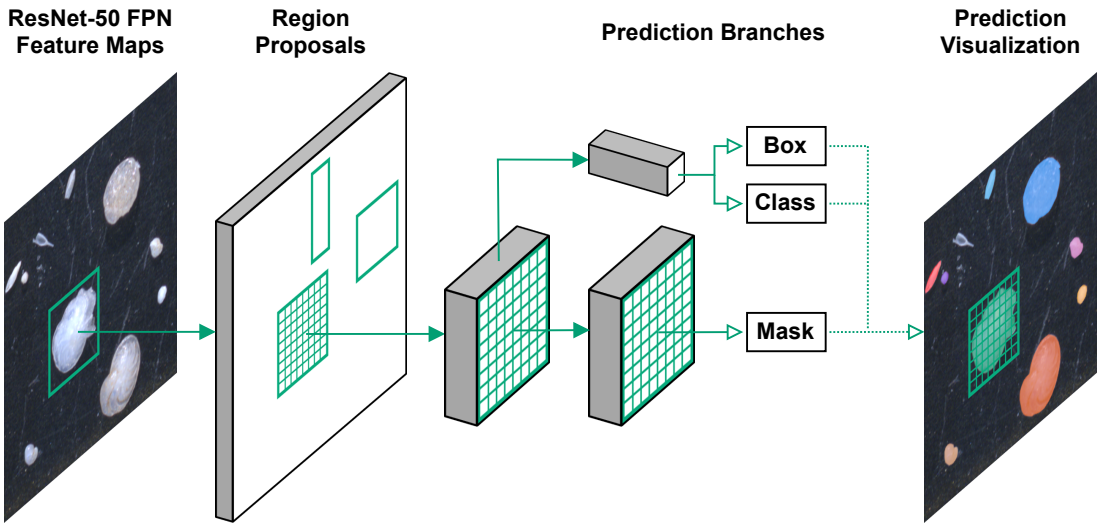


Figure 3.10. *Sketch-like depiction of the Mask R-CNN architecture.*

Mask R-CNN [19] is a relatively recent deep learning model that can perform instance segmentation in a single forward-pass, and is an extension of the Faster R-CNN [80] model. Faster R-CNN is an object detection model, meaning it predicts bounding boxes and class labels, and this is performed in two-stages during a single forward-pass. The first stage produces candidate region proposals via feature maps extracted from a so-called backbone, which is generally a feature pyramid network [81]. The candidate region proposals are refined by an attention-like network called a region proposal network (RPN). Additionally, the number of region proposals is reduced by performing non-maximum suppression based on intersection-over-union thresholds and class scores. The second stage is detection via a Fast R-CNN [82] model, which ultimately predicts bounding boxes and class labels for a set of proposed region of interests (RoIs). Mask R-CNN extends the Faster R-CNN architecture in two key aspects; first it adds a decoupled segmentation mask prediction branch, implemented as a small fully convolutional network (FCN) [83]. The mask branch predicts class-specific binary segmentation masks for each RoI, which is unlike typical FCN-based segmentation models that predict multinomial masks. Second, it replaces the RoI pooling layer of Fast R-CNN with a new "RoIAlign" layer, which preserves the exact spatial locations of each RoI, which is necessary for predicting good segmentation masks [19]. Figure 3.10 depicts a high-level summary of the Mask R-CNN architecture.

The Mask R-CNN architecture is very flexible and allows a wide-variety of backbone models to be used, which further allows e.g. accuracy to be traded for computational speed. However, several comparable architectures have been proposed that report both increased accuracy

and computational speed-up, as well as other improvements. Some examples of these models include PANet [84], TensorMask [85], and CenterMask [86]. It is also worth noting that Mask R-CNN and its derivatives are two-stage models, meaning they first perform a detection step, followed by a joint bounding box, classification and segmentation step; but recently, single-stage models have surpassed many two-stage models such as Mask R-CNN both in terms of accuracy and speed [87]. This is a promising advancement compared to the relatively high computational cost of Mask R-CNN, which makes it unsuitable for many real-time applications.

3.6. Transfer learning

The idea behind transfer learning is that knowledge acquired when learning to perform one conceptual task, should make it easier to learn other, similar, conceptual tasks [84]. We can intuitively think of this in the context of a real-world example; learning to play musical instruments. It is not unreasonable to assume that it will be easier for someone to learn to play a new instrument if they have already learned how to play one or more other instruments. While the instruments themselves might be very different, we can assume that some conceptual knowledge of playing music, e.g. reading notes and sheet music, will be largely the same. We might consider that the task is the same (playing an instrument), but each instrument has their own domain where the feature space varies from instrument to instrument. In general, the more similar two domains are, the more knowledge can be transferred from one to the other [88].

We can define a source domain D_S as the domain in which the model has been trained, where we can think of the domain as the origin of the training data. The target domain D_T is then defined as the domain we want to transfer a model towards, assuming the model has already been trained on a source domain. Additionally, each domain has its own domain data, meaning that we have a source input feature space $\mathcal{X}_S$ and a target input feature space $\mathcal{X}_T$. In a supervised learning setting we also have a source label space $\mathcal{Y}_S$ and target label space $\mathcal{Y}_T$. Similarly, on the task level, we define the source task $\mathcal{T}_S$ as the learning task the model was trained to perform; e.g. predicting class labels on the ImageNet dataset. The target task $\mathcal{T}_T$ is the new task we want to transfer the source task knowledge to; e.g. predicting microscopic foraminifera class labels using a model trained the ImageNet dataset. When both label spaces, $\mathcal{Y}_S$ and $\mathcal{Y}_T$, are known, the transfer learning problem is referred to as transductive. If only the source label space $\mathcal{Y}_S$ is known, the problem can be categorized as inductive transfer learning. When neither label space is known, we categorize the situation as an unsupervised transfer

learning problem [89]. Transfer learning can also be categorized based on the similarity between the feature and label spaces. Homogeneous transfer learning is defined as the setting where $\mathcal{X}_S = \mathcal{X}_T$ and $\mathcal{Y}_S = \mathcal{Y}_T$. Conversely, heterogeneous transfer learning is defined as the setting where $\mathcal{X}_S \neq \mathcal{X}_T$ or $\mathcal{Y}_S \neq \mathcal{Y}_T$ [90].

Transfer learning in the context of machine learning is widely used in many applications, and is especially useful in domains where it is very costly, difficult, or otherwise infeasible to gather the large amount of (labeled) data frequently needed to train complex deep learning models from scratch [90]. As a field of research within machine learning, transfer learning is not new, but it has become increasingly relevant since the deep learning revolution in the last decade. The field has seen many algorithmic developments, as well as theoretical developments, in recent years [84]. For many machine learning tasks we can leverage pre-trained machine learning models trained on one source domain and transferring to a new target domain. This is common in areas such as computer vision where state-of-the-art deep learning models are typically being trained on datasets containing millions of labeled images [17, 91]. Collecting and labeling e.g. millions of images, and subsequently training complex models on such large datasets, is something that is not accessible to most researchers. However, with the advancements in transfer learning, it is possible to learn models on much smaller datasets and still get good task performance measures [92–95].

There are many advanced transfer learning methods with well-established theoretical frameworks, but perhaps the most widely used technique, referred to as fine-tuning, is remarkably simple. The first step in fine-tuning is to find a model that has been trained on a source domain that has a feature space that is equivalent to the feature space in the target domain. The underlying idea behind fine-tuning is that the pretrained model has learned some general feature extractors that are transferable from the source domain to the target domain, and that with some degree of parameter fine-tuning the model can applied to the target domain. Typically, the amount of model parameter tuning required depends on how closely aligned the domains and feature spaces are.

There are several ways to implement fine-tuning, but let us consider an approach often applied in image classification with CNN models. It is very common to use a CNN classifier pretrained on the ImageNet [18] dataset, with the assumption that images in the target domain are equivalent to the source domain. The key insight here is that the first few layers of most deep CNN classifiers are general-purpose image feature extractors that are transferable to many image domains [96]. Typically, it is only the few last layers in the model, which form the higher-level feature maps that tend to be more domain specific, and it is therefore only these layers that are fine-tuned. The classification head of the pretrained model, which is

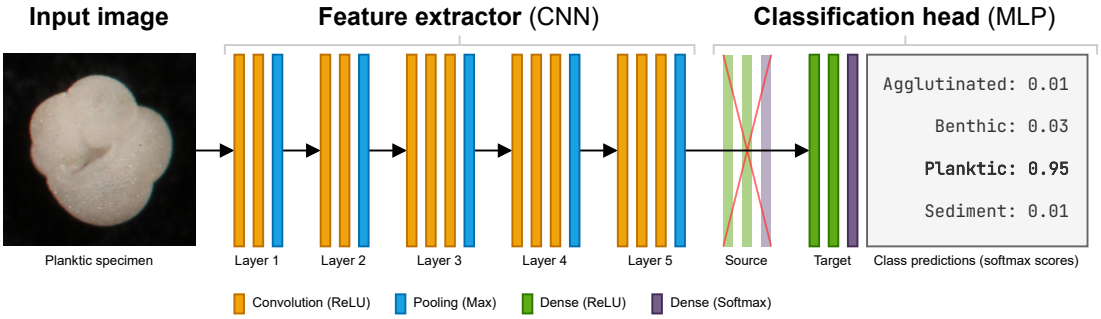


Figure 3.11. *Illustration of fine-tuning a pretrained classifier. The original classification head is replaced with a new head that is designed for the desired classification task. Input to the new classification head are the features from the pretrained feature extractor. Frequently, layers of the feature extractor are fine-tuned on the new data to produce better features, and often it is sufficient to fine-tune the last few layers. Note that the layer term here is sometimes also referred to as a "block."*

domain specific, is replaced with a new classification head designed for the target domain. See Figure 3.11 for an example of this approach.

There are many ways to perform the actual model fine-tuning of the new target domain model, but let us consider two variations that are common in many practical applications. The first approach involves initially keeping the parameters of the pretrained feature extractor unchanged, and only learning the parameters of the new classification head. After some initial training of the classification head, the parameters of the pretrained feature extractor are then fine-tuned together with the parameters of the classification head. The second approach is similar to the first one, but the initial pretraining of the new classification head is skipped, and the parameters of the entire model are adapted and learned, end-to-end, from the beginning. This approach can be beneficial when the source and target feature spaces are not very consistent, such that features extracted from the pretrained model are not sufficient for training the new classification head.

Transfer learning is not without challenges however, where two concrete examples are negative transfer [88] and catastrophic forgetting [97]. Negative transfer occurs when knowledge learned about the source domain and source task are detrimental to the performance of the target task in the target domain. Reportedly, the challenge of negative transfer still remains largely unsolved, but some progress has been made [98]. The challenge of catastrophic forgetting is described as the knowledge learned about the source domain and task are "forgotten" when a model is transferred to new targets. We can relate this to the fine-tuning methodology; when adapting the model parameters to a new target domain, as the model performance

on the target task increases, performance on the source task is reduced. Intuitively, this occurs because the parameters learned by the feature extractor in the source domain are being changed, even if ever so slightly, and this impacts the performance unless regularized. Techniques such as Learning without Forgetting [97] addresses catastrophic forgetting by ensuring that source domain predictions remain unchanged while transferring to the target domain.

3.7. Uncertainty estimation

Uncertainty in deep learning is valuable because it can give an insight into what a model does not know, and when we can trust the predictions it produces. In general we can divide uncertainty into two types; aleatoric uncertainty and epistemic uncertainty. Aleatoric uncertainty is caused by noise inherent in the data or the process that produced the data, and it can not be reduced by introducing more data — even in the limit of infinite data — and is therefore sometimes referred to as irreducible uncertainty [99]. Two examples of aleatoric uncertainty are images with motion blur (smoothing) caused by movement, and measurement noise caused by imperfections in the acquisition sensors. Epistemic uncertainty originates from the uncertainty in the model parameters, and can be thought of as lack of knowledge, or more specifically, a lack of data [100]. This type of uncertainty is generally high for out-of-distribution data. Epistemic uncertainty is reduced by training on more data and (under the assumption of identifiability) vanishes in the limit of infinite data. In safety-critical applications, such as computer vision for autonomous vehicles and computer-aided decision systems in healthcare, epistemic uncertainty is very important. Knowing when the model is uncertain, and incorporating that into relevant decision processes, can avoid disastrous outcomes [99].

Bayesian modeling, and more specifically Bayesian neural networks, can be considered the optimal way to model both aleatoric and epistemic uncertainty in most applications. In these networks, deterministic point estimates of model parameters are replaced with prior distributions over the weights. And direct optimization of model parameters is replaced by marginalization over all possible weights. Additionally, due to intractability of evaluating posteriors in Bayesian neural networks analytically, approximate inference techniques are employed instead [99]. Although Bayesian neural networks have many benefits, they are still challenging to train and generally have much higher computational costs when compared to other deep learning methods. Therefore, when real-time application is vital, or when training very complex deep learning models, Bayesian approximation methods are often preferred.

A popular technique for estimating the epistemic uncertainty of deep learning models is the approach usually referred to as Monte Carlo dropout [101]. The method is based on the realization that dropout can be considered a Bayesian approximation, in the sense that every neural network with randomly "dropped" neurons is approximately equivalent to sampling a network from a distribution of networks. By extension, this then opens up the door to computing a Monte Carlo approximation of the model predictive uncertainty. Importantly, Monte Carlo dropout is remarkably simple to implement and can be applied to any model with dropout layers. Assume we have a neural network f with parameters $\boldsymbol{w}$ such that

$$\hat{\boldsymbol{y}} = f(\boldsymbol{x}; \boldsymbol{w}), \tag{3.16}$$

where $\hat{\boldsymbol{y}}$ are predicted outputs for the input features $\boldsymbol{x}$ with true outputs $\boldsymbol{y}$. If we collect k Monte Carlo samples of the model predictions,

$$\hat{\boldsymbol{y}}_i = f(\boldsymbol{x}; \boldsymbol{w}_i), \quad i = 1, 2, \ldots, k, \tag{3.17}$$

where $\boldsymbol{w}_i$ are the model parameters the i-th Monte Carlo sample after applying dropout. The model predictive uncertainty and variance can then be calculated [101],

$$\hat{\boldsymbol{\mu}} = \frac{1}{k} \sum_{i=1}^{k} \hat{\boldsymbol{y}}_i, \tag{3.18}$$

$$\hat{\boldsymbol{\sigma}} = \frac{1}{k} \sum_{i=1} (\hat{\boldsymbol{y}}_i - \hat{\boldsymbol{\mu}})^2. \tag{3.19}$$

One drawback with Monte Carlo dropout is that it requires the use of dropout layers, but many deep learning models that rely on batch normalization [102] layers are not using dropout. This means that Monte Carlo dropout is not available when using any of these models, and adding dropout to models after-the-fact can have detrimental impact on task performance. Recently, a variation of Monte Carlo dropout has been proposed, which is referred to as Monte Carlo batch normalization [103]. Just as Monte Carlo dropout can be applied to any network with dropout layers, Monte Carlo batch normalization can be applied to any network with batch normalization layers. The key insight behind the method is the realization that during training, the statistical moments of each batch normalization layer are updated for each batch, and are thereby stochastic if the batches are stochastic. During inference the moments are normally "fixed" and deterministic, using parameters based on the running statistics computed during training. However, by running inference with batch normalization layers in "training mode", stochasticity is introduced by preceding all target

predictions with random batches of data. This introduces an equivalent level of randomness to the model akin to Monte Carlo dropout, which results in the configuration of the neural network being stochastic as if sampled. The procedures for estimating the model predictive uncertainty and ensemble predictions are the same as for Monte Carlo dropout.

3.8. Scale-space techniques

Scale-space theory is a framework for multi-scale representation, which was originally developed for applications in computer vision [104]. The underlying idea of scale-space techniques is that in many cases objects or features are only meaningful at certain scales, and often these scales are unknown. Another, perhaps slightly more intuitive way to think about this is that at large scales, most small features are essentially equivalent with noise. Conversely, at small scales it does not make sense to consider comparisons with large-scale features. This makes it challenging when developing automated systems for interpreting or analyzing data, e.g. change detection in images [105]. In the scale-space framework this is resolved by considering scale-space representations at all scales concurrently, where fine-scale structures are gradually removed. Scale-space representations of images are typically produced by convolving the input with Gaussian kernels or Gaussian kernel derivatives [104].

Since its inception in computer vision, scale-space theory has been extended to statistical analysis, where it was first applied to mode detection for uni- and bivariate density estimation [106]. In the original scale-space methodology concerning images, the scale-space representations are often referred to as a family of “blurs”; similarly, for statistical scale-space, where the application is primarily curves, the scale-space representations are said to be a family of smooth functions — or “smooths”. One of the most important methods in statistical scale-space analysis is the SiZer methodology by Chaudhuri and Marron [107]. The SiZer idea is based on non-parametric curve fitting and answering the question: which features are “real” features, and which are simply random noise in the data. This question is addressed by statistical inference, via significance/credibility testing of the scale-dependent features, and the results are typically presented in easily interpretable visualizations referred to as “maps” [106]. An important point in SiZer is that focus is shifted from the search for an underlying true signal to an analysis of scale-space versions of the unknown signal. By this approach, two immediate advantages can be utilized. Firstly, the scale-space idea avoids bias problems since the scale-space signals are unbiased estimates of the true smooth signals. Secondly, since all bandwidths are relevant, the search for an optimal smoothing can be dispensed with.

The SiZer methodology has turned out to be very useful in many applications and it has been extended beyond one-dimensional curve fitting, such as a two-dimensional variant applied to images [108, 109]. It has also seen the inclusion of Bayesian modeling with the BSiZer [110] variant, which also includes image-based variant called iBSiZer [105]. BSiZer has been successfully applied to research fields such as paleoclimate reconstruction and climate prediction [111].

Ideas for applying scale-space methodology in hyperspectral imaging is given in “Scale-space in hyperspectral image analysis” [112]. A first attempt to extend the SiZer methodology to hyperspectral data so that changes in images can be detected at a very early stage is given in Paper II of this dissertation. Since it is very hard to know what kind of changes that may occur in such images, a scale-space approach makes sense.

Part II.

Summary of research

Chapter 4.

Paper I — Recent advances in hyperspectral imaging for melanoma detection

This paper is a critical review of peer-reviewed research in the intersection of machine learning and skin cancer detection/classification using hyperspectral images in the years from 2003 to 2018. We conducted this research in order to identify what had been done, and to identify gaps and possible research ideas towards detecting melanoma skin cancer using hyperspectral images. Our hypothesis was that not much research had been conducted in this field, and that most published work would be based on a multitude of different imaging systems, datasets, and so forth. The initial literature search uncovered 86 candidate publications, which were reduced to 20 after applying exclusion criteria for relevance, and vetting the quality of the published research. The 20 included papers were then thoroughly analyzed and synthesized. This process included investigating the methodology, imaging systems, datasets, results, conclusions, etc. The review presents our thorough analysis and critical remarks, as well as future research directions, both in the short- and long-term. Finally, we provide a short list of recommendations for reducing the risk of repeating mistakes highlighted in the review, in order to ensure valid and reliable results in future work.

Contributions by the author:

- Defined the initial scope, search strategy, and inclusion/exclusion criteria.
- Conducted the initial literature search, and reduced the list of candidate publications by applying exclusion criteria.
- Defined and organized the research synthesis workflow used by all contributors.
- Summarized the key findings uncovered during the research synthesis phase.

- Wrote the initial draft, and was in charge of producing the final manuscript.

Chapter 5.

Paper II — Early Detection of Change by Applying Scale-Space Methodology to Hyperspectral Images

The work in this paper is based on the idea of detecting small changes in the reflectance curves in hyperspectral images. Early change detection is deemed important in the case of skin cancer to reduce the chances of the cancer going undetected until it becomes metastatic and begins spreading to other organs, which drastically reduces the survival rate. The initial plan was to capture multiple hyperspectral images of the same skin lesion over time in order to monitor it. But due to the fact that any suspicious skin lesion is surgically removed, this turned out to not be feasible within the constraints of the project. Therefore, we instead introduced minor changes to a few spectral signatures in real hyperspectral skin lesion images to simulate a time-based evolution. Change detection is also important in applications such as automated food quality analysis, and here we were able to acquire several hyperspectral images of frozen fish. These images had been acquired at several different time steps in order to monitor the quality of the fish over time. We propose a scale-space methodology suitable for both applications, and present results that show the methodology is indeed capable of detecting changes in the spectral signatures. Finally, we discuss challenges and opportunities, as well as promising directions for future research.

Contributions by the author:

- Sourced a prototype hyperspectral image acquisition system for clinical dermatology applications.
- Acquired, preprocessed and prepared a novel dataset of hyperspectral skin lesion images with histopathology verified diagnostic labels.

- Prepared the hyperspectral images of the frozen fish for further analysis.
- Involved in the conceptualization of the scale-space change detection methodology.
- Participated in the writing of the manuscript—reviewing and editing.

Chapter 6.

Paper III — Towards detection and classification of microscopic foraminifera using transfer learning

This paper is the result of what started as a proof of concept for automatically classifying microscopic foraminifera using computer vision. In this paper we first acquired images of microscopic foraminifera and sediment grains using a microscope-based acquisition system to produce a novel dataset. Each image consisted of a large number of candidate objects, which after several preprocessing steps resulted in a dataset containing approximately 2600 labeled images of individual objects divided into four high-level classes; agglutinated benthics, calcareous benthics, planktics, and sediment grains. The classification methodology used in the paper is based on implementing a classifier based on a VGG-16 model with parameters pretrained on the ImageNet dataset, and then training and fine-tuning the full set of model parameters on the foraminifera dataset. Additionally, to improve the usefulness in a real-world situation we used the so-called Monte Carlo dropout technique to estimate the predictive uncertainty of the model. We leverage the uncertainty to uncover insightful negative predictions and out-of-distribution examples from the test dataset, e.g. images of overexposed objects, or objects missing their characteristic morphology due to their orientation under the microscope. We present classification accuracy using classifiers with and without Monte Carlo dropout sampling, as well as classification accuracy using an ensemble-based approach with a majority vote scheme. The results are very promising for the high-level class labels, with a classification accuracy of $98.8 \pm 0.2\%$ for 10 independently trained models without dropout sampling. Using the Monte Carlo dropout implementation with 100 Monte Carlo samples, we got a mean accuracy of $97.9 \pm 0.5\%$. The implementation of ensemble predictions with a majority-vote scheme yielded an accuracy of 98.5% Finally, we provide

some concluding remarks with suggested future improvements.

Contributions by the author:

- Preprocessed and curated a novel dataset of more than 2600 labeled images of microscopic foraminifera.
- Implemented the VGG-16 based classifier, and the Monte Carlo dropout sampling.
- Designed and ran all experiments.
- Performed the analysis and synthesis of results, which included making all figures in the manuscript.
- Wrote the initial draft of the manuscript, as well as reviewed and edited the manuscript.
- Prepared the final version of the manuscript.

Chapter 7.

Paper IV — Instance Segmentation of Microscopic Foraminifera

This paper is based on instance segmentation of microscopic foraminifera and deep learning. The work is an extension of the work started in paper III, but includes two new sources of datasets. In the paper we first introduce a novel object detection dataset, based on three different rounds of image acquisition and two different microscope setups. The dataset contains 104 images containing over 7000 objects with corresponding bounding boxes and segmentation masks. We then present an instance segmentation model, based on Mask R-CNN that has been pretrained on the COCO dataset, which we subsequently fine-tuned on our novel dataset. Our results show a COCO-style average precision (AP) of 0.78 for the bounding box regression task and 0.80 for the segmentation mask prediction task. The average recall (AR) for both tasks is 0.83 and 0.84, respectively. We also investigated the model performance on a per-class level, and based on this discovered challenges with several images containing dense clusters of objects from the "sediment" class. By excluding this class from the model evaluation, the AP for the bounding box task increased to 0.84, and to 0.86 the segmentation mask task. When we evaluated the model with different IoU thresholds, we found that near pixel-perfect predictions is challenging for the model. Significant increases in AP and AR were gained for both the bounding box and segmentation tasks for IoU thresholds below 0.95. Our qualitative analysis of the trained instance segmentation model correspond well with the reported precision and recall scores. We end the paper with a discussion of our findings, and propose a list of several key directions for future research.

Contributions by the author:

- Preprocessed and curated a novel object detection dataset of more than 7000 objects with high-quality segmentation masks and class labels.

- Implemented and adapted an instance segmentation model based on the Mask R-CNN architecture.
- Designed and ran all experiments.
- Performed the analysis and synthesis of results, which included all figures and visualizations of predicted segmentation masks and detections.
- Wrote the initial draft of the manuscript.
- Reviewed and edited the manuscript together with co-authors.
- Prepared the final version of the manuscript together with the third author.

Chapter 8.

Concluding remarks

In this dissertation we have leveraged computer vision methodologies to address challenges and opportunities in biomedicine and geoscience. In the biomedicine field, we have proposed a novel scale-space methodology towards the early detection of change in hyperspectral images of skin lesions. The method is designed to detect changes in spectral signatures over time, which can be an important bio-indicator of skin cancer. We also demonstrated the effectiveness of the method on hyperspectral images of other biological, time-varying targets. This work, as well as our critical review of recent advancements in hyperspectral imaging for melanoma detection, are important contributions towards early detection of melanoma skin cancer.

In the geoscience field, we have proposed two methodologies towards automating the identification, counting, and picking of microscopic foraminifera. First, we developed a deep learning-based classification model, which was trained on a novel dataset of more than 2600 labeled images of microscopic foraminifera and sediment grains. The classification model achieved a very high accuracy, and verified its robustness by implementing a Monte Carlo sampling approach for estimating the model predictive uncertainty. Second, we created a novel object detection dataset of more than 7000 microscopic foraminifera and sediment grains, which were used to train a deep learning-based instance segmentation model. We demonstrated that the instance segmentation model achieved high accuracy both at detecting objects and delineating them with fine-grained segmentation masks. Both deep learning methodologies are valuable contributions for automating the process of identifying, counting, and picking microscopic foraminifera.

In conclusion, we have addressed the four opportunities presented in the introduction of this dissertation by the contributions presented in the four included papers.

8.1. Limitations and future work

In Paper I, the biggest limitation lies in its scope; we limited the study to research focused on detecting/classifying skin cancer using multi- and hyperspectral images. Including other skin lesions and dermatological conditions, would have drastically increased the scope, but would likely have identified many other important opportunities. As with all literature reviews, they have an expiry date, and in the time since our work was published many important works have been published. Therefore, in the future it would be worth considering publishing an updated review, which takes into account all recent developments.

When we started working on Paper II, our intent was to monitor cases of skin cancer over time, but this turned out to not be possible because any suspected cancer is surgically removed. Therefore, we had to simulate an evolution in the hyperspectral images of melanoma skin cancer by introducing very small changes in the spectral signatures. While this was a reasonable and pragmatic approach, it is not clear-cut if this is directly transferable to a real-world application. Possible future work in this direction could be a shift away from following melanoma skin cancer over time, to other skin lesions that typically evolve. In particular, it would be valuable to monitor precancerous skin lesions over time until they evolve into suspected skin cancers and are removed.

In Paper III and IV, the biggest limitation is caused by the class labels in both datasets, which we limited to four high-level categories. This was done primarily due to limited resources; accurately labeling each object is very time-consuming. Introducing many more fine-grained, species-level class labels also would have required collecting a lot more training data. Identifying and counting the foraminifera species is important in most applications, so future work should be focused on expanding the datasets with better labels. Additionally, in Paper IV we wanted to implement uncertainty estimation for the segmentation masks, but due to time constraints this was abandoned. This should be addressed in future work, and we suggest that it can be solved based on work presented in ...

Part III.

Included papers

Chapter 9.

Paper I

Recent advances in hyperspectral imaging for melanoma detection

Thomas Haugland Johansen, Kajsa Møllersen, Samuel Ortega, Himar Fabelo, Aday Garcia, Gustavo M. Callico, Fred Godtliebsen

Published

Received: 7 December 2018 | Revised: 14 March 2019 | Accepted: 18 March 2019

DOI: 10.1002/wics.1465

ADVANCED REVIEW

WIREs COMPUTATIONAL STATISTICS WILEY

Recent advances in hyperspectral imaging for melanoma detection

Thomas Haugland Johansen[1] | **Kajsa Møllersen**[2] | **Samuel Ortega**[3] | **Himar Fabelo**[3] | **Aday Garcia**[3] | **Gustavo M. Callico**[3] | **Fred Godtliebsen**[1]

[1]Department of Mathematics and Statistics, UiT The Arctic University of Norway, Tromsø, Norway

[2]Department of Community Medicine, UiT The Arctic University of Norway, Tromsø, Norway

[3]Institute for Applied Microelectronics, University of Las Palmas de Gran Canaria, Las Palmas, Spain

Correspondence
Thomas Haugland Johansen, Department of Mathematics and Statistics, UiT The Arctic University of Norway, Tromsø, Norway.
Email: thomas.h.johansen@uit.no

Funding information
Agencia Canaria de Investigación, Innovación y Sociedad de la Información (ACIISI), Grant/Award Number: ProID2017010164; Spanish Government and European Union (FEDER funds), Grant/Award Number: TEC2017-86722-C4-1-R; European Social Fund (FSE) and Agencia Canaria de Investigacion, Innovacion y Sociedad de la Información (ACIISI), Grant/Award Number: POC 2014–2020, Eje 3 Tema Prioritario 74(85%); Tromsø Forskningsstiftelse, Grant/Award Number: A33020

Abstract

Skin cancer is one of the most common types of cancer. Skin cancers are classified as nonmelanoma and melanoma, with the first type being the most frequent and the second type being the most deadly. The key to effective treatment of skin cancer is early detection. With the recent increase of computational power, the number of algorithms to detect and classify skin lesions has increased. The overall verdict on systems based on clinical and dermoscopic images captured with conventional RGB (red, green, and blue) cameras is that they do not outperform dermatologists. Computer-based systems based on conventional RGB images seem to have reached an upper limit in their performance, while emerging technologies such as hyperspectral and multispectral imaging might possibly improve the results. These types of images can explore spectral regions beyond the human eye capabilities. Feature selection and dimensionality reduction are crucial parts of extracting salient information from this type of data. It is necessary to extend current classification methodologies to use all of the spatiospectral information, and deep learning models should be explored since they are capable of learning robust feature detectors from data. There is a lack of large, high-quality datasets of hyperspectral skin lesion images, and there is a need for tools that can aid with monitoring the evolution of skin lesions over time. To understand the rich information contained in hyperspectral images, further research using data science and statistical methodologies, such as functional data analysis, scale-space theory, machine learning, and so on, are essential.

This article is categorized under:

Applications of Computational Statistics > Health and Medical Data/Informatics

KEYWORDS

hyperspectral, machine learning, melanoma, skin cancer

1 | INTRODUCTION

Skin cancer is one of the most common types of cancer in humans, and in countries with predominantly fair-skinned population, the incidence trend for the past 30 years has been increasing (American Cancer Society, 2018; Ferlay et al., 2013). Skin

cancers are classified as nonmelanoma skin cancer (NMSC) and melanoma. NMSC is by far the most frequent, whereas melanoma is the most deadly. In 2018, the reported number of new cases of NMSC globally accounted for 5.8% of all new cancer cases, and accounting for 0.7% of all deaths. New cases of melanoma was reported to account for 1.6% of new cancer cases, but notably accounting for 0.6% of all deaths caused by cancer (Bray et al., 2018). The key to effective treatment of skin cancer is early detection, before the cancer metastasizes. Nonmetastasized melanoma is reported to have a 5-year survival rate of 99%, whereas once it spreads to distant organs the survival rate drops to 20% (American Cancer Society, 2018). In dermatology, one of the most commonly taught diagnostic guidelines for classifying pigmented skin lesions is the ABCD rule of dermatoscopy (Nachbar et al., 1994). The respective letters in the acronym represent different features of a skin lesion: asymmetry, border, color, and differential structures. When using the ABCD rule to diagnose a skin lesion, a score is assigned for each of the four features, and combined into a total score. The total score gives an indication of the potential for malignancy, where higher scores mean greater potential for malignancy. In clinical settings, the reported sensitivity and specificity of the ABCD rule ranges from 74 to 91.6% and 45 to 67%, respectively (Ahnlide, Bjellerup, Nilsson, & Nielsen, 2016; Annessi, Bono, Sampogna, Faraggiana, & Abeni, 2007; Unlu, Akay, & Erdem, 2014). Figure 1 shows a few examples illustrating typical variation between different skin lesions. Note the differences in the shapes, borders (and lack thereof), colors, and so on.

Given the increasing trend in skin cancer prevalence, and the difficulty in detecting skin cancer at an early stage, researchers across many fields have been working to both extend and develop new diagnostic criteria and computational algorithms. For example, the ABCD rule of dermoscopy has been extended to ABCDE, where the E accounts for evolution of the skin lesion over time (Abbasi et al., 2004). With the advent of machine learning and the increasing access to vast, inexpensive computational power, several research groups have been focusing on developing automated and semiautomated computational methods for detecting and classifying skin lesions. While some recent advances have been developed using conventional RGB (red, green, and blue) imaging techniques (Esteva et al., 2017), other researchers have been focusing on exploring new avenues of skin cancer classification using multispectral and hyperspectral imaging techniques.

Computer systems for classification of pigmented skin lesions have been an active research field for several decades. Early systems used conventional RGB images, but by the early 2000s almost all systems used dermoscopic images (Rosado et al., 2003). See Figure 2 for examples of both types of images. A dermoscope is a simple device consisting of a magnifying lens, a glass plate, and a light source that allows the light to penetrate the uppermost layer of the skin. It is commonly used by dermatologists. The overall verdict of systems based on conventional and dermoscopic images is that they do not outperform

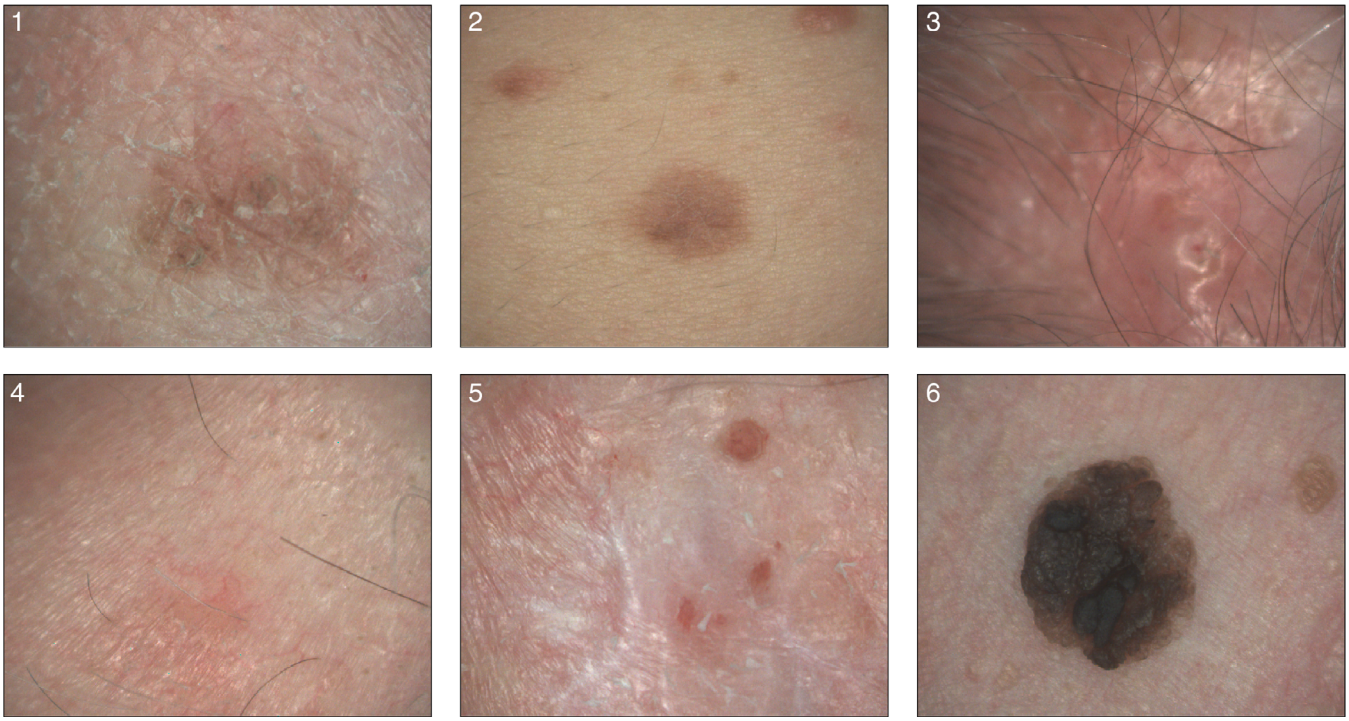


FIGURE 1 Examples of melanoma and nonmelanoma skin cancer taken in a clinical setting from six different patients. These cases represent both nonmelanoma and melanoma skin cancer. The diagnoses of the lesions based on histopathology are as follows: (1) melanoma, (2) atypical melanocytic hyperplasia, (3) squamous cell carcinoma, (4) Bowen's disease, (5) basal cell carcinoma, and (6) seborrheic keratosis

dermatologists (Korotkov & Garcia, 2012; Rosado et al., 2003; Vestergaard & Menzies, 2008). Deep learning has been introduced to skin lesion classification, and although deep learning methods possibly outperform traditional approaches (Codella et al., 2018), it has not outperformed the dermatologist (Esteva et al., 2017). Multispectral imaging increases the amount of retrieved information and various systems have been used for skin lesion classification. Whether this increases the performance has not been established with certainty, since dermoscopic and multispectral systems have not been tested on the same set of lesions, or under strictly similar conditions. The conventional and dermoscopic systems seem to have reached an upper limit for their performance, while emerging technologies such as hyperspectral imaging can possibly increase the performance.

The main advantage of hyperspectral and multispectral imaging compared with conventional imaging technologies is the possibility of exploring spectral regions beyond the human eye capabilities. Some materials present spectral features in the infrared spectral range (Lachenal & Ozaki, 1999). Besides the spectral range, the use of hyperspectral images is necessary when the material being analyzed presents narrow spectral features (Jet Propulsion Laboratory, California Institute of Technology, n.d.; Lee, Cohen, Kennedy, Maiersperger, & Gower, 2004). Such narrow spectral features cannot be detected using multispectral or RGB images, and should therefore be measured using high spectral resolution instrumentation. Figure 3 illustrates the difference in fidelity and richness of hyperspectral images in comparison with conventional RGB images.

In this review, we will report on the recent advances that specifically focus on detecting skin cancer using multi- and hyperspectral images. We will start by giving a short description of the review methodology. Then, we give a brief introduction to hyperspectral imaging and point out how this imaging technique is being used in medicine, and specifically why it is being used to classify skin cancer. Next, we focus on how feature selection is crucial for extracting information of this type of data and thereafter we point out the need for extending current classification methodologies to include the use of spatiospectral information. Our review finally gives some critical remarks and analysis of relevant published results before we indicate important future research directions.

2 | REVIEW METHODOLOGY

The goal of this review was to provide insight into recent advances in detection of skin cancer using hyperspectral imaging systems in order to uncover what has been achieved, and to understand what the key challenges are. Based on this we defined

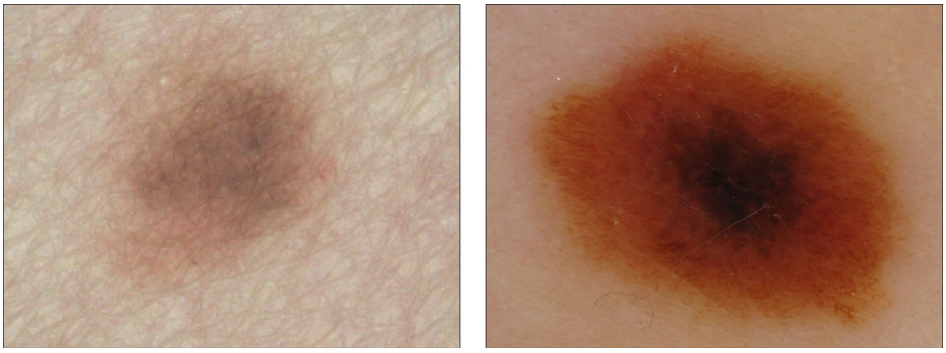


FIGURE 2 The image on the left is an example of a conventional, clinical image of a pigmented skin lesion, whereas the image on the right is an example of dermoscopic image

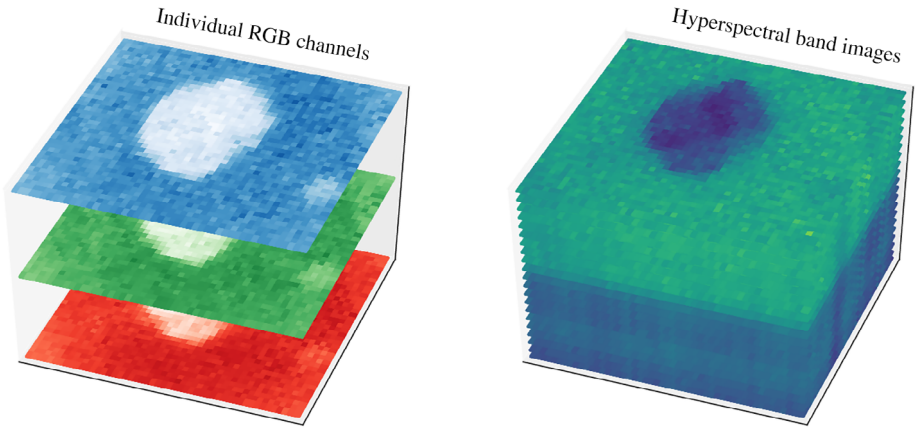


FIGURE 3 The conceptual difference between the information richness in a hyperspectral cube and an RGB image. In the hyperspectral cube, each horizontal slice represents spatial response for a discrete wavelength. For the RGB image, each slice represents spatial information across a range of wavelengths. Each of the red, green, and blue slices are calculated based on the visual light spectrum associated with each respective color

the following inclusion criteria with the intention of only including recent, highly relevant, peer-reviewed publications focusing on skin cancer detection using hyperspectral images;

- Peer-reviewed publication in journal or conference proceeding.
- Based on hyperspectral (or multispectral) images.
- Specifically dealing with skin lesion classification.
- Noninvasive data collection, that is, in vivo skin lesions.
- Published in recent years (2003–2018).

The inclusion of multispectral imaging systems was made based on preliminary searches, which uncovered that most of the relevant skin cancer research has been done with these systems. Although multispectral and hyperspectral systems are based on different concepts and technologies, from the perspective of data analysis and pattern recognition, images produced by these systems present similar benefits and challenges. Our initial threshold for "recent" was 10 years, but because some very relevant studies were published more than 10 years ago, the threshold was increased to 15 years.

In the period of August 20–23, 2018, we performed searches on Web of Science, PubMed, Scopus, and Google Scholar. Search queries were specifically adapted to each search engine, and based on the search criteria seen in Listing 1.

Listing 1: The search criteria used to construct search queries for Web of Science, PubMed, Scopus, and Google Scholar.

```
((multispectral AND classification) OR hyperspectral) AND
(image OR camera) AND
(skin OR melanoma) AND
(classification OR lesion OR cancer)
```

Our searches resulted in a collection of 86 peer-reviewed publications that were potential candidates for review based on their titles, keywords, and abstracts. After reading through the initial collection of candidates, we ended up with 20 publications relevant for the review, selected based upon our previously listed inclusion criteria. Figure 4 shows a breakdown of the number of publications per year, and a summary of all the reviewed publications can be seen in Table 1.

3 | HYPERSPECTRAL IMAGING FOR SKIN CANCER CLASSIFICATION

Hyperspectral imaging has shown considerable potential as a noninvasive and nonionizing technique, supporting rapid acquisition and analysis of diagnostic information. Unlike conventional RGB cameras, which are limited to capturing three bands in the electromagnetic spectrum, hyperspectral imaging systems are capable of capturing hundreds of narrow bands across the electromagnetic spectrum, both inside and outside the human visual spectral range (Smith, 2012). Hyperspectral imaging has been widely used in remote sensing (Tuia, Volpi, Copa, Kanevski, & Munoz-Mari, 2011), and has been applied in vitro, ex vivo, and in vivo in different medical applications (Lu & Fei, 2014). For skin lesion classification, several studies have been conducted using different types of hyperspectral and multispectral acquisition systems. Based on the reviewed publications listed in Table 1, most of the research effort up until now has been based on multispectral systems. Some multispectral devices for skin lesion analysis are commercially available, such as MelaFind (Elbaum et al., 2001; Kupetsky & Ferris, 2013)

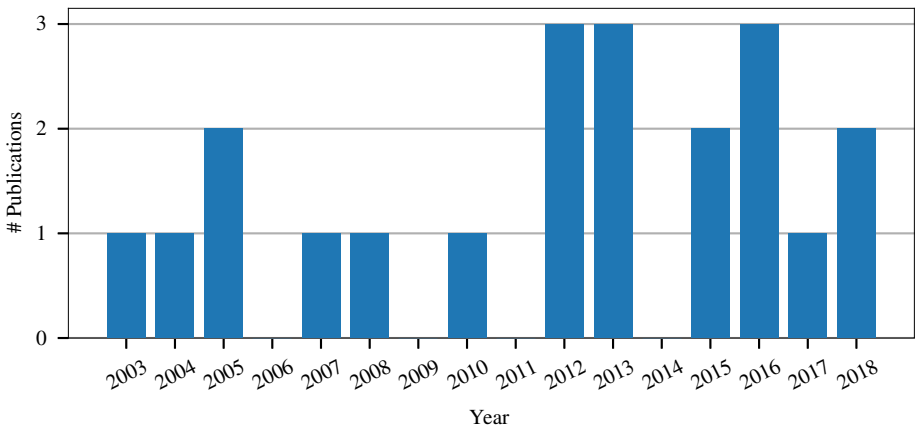


FIGURE 4 The number of publications per year that matched our search queries and were selected for review based on our inclusion criteria. From the plot, we can see that the majority of reviewed publications were published after 2011

TABLE 1 Summary of the publications included in the review

Publication	Imaging system	Wavelengths (nm)	Bands	Pixel count
Tomatis et al. (2003)	Custom	400–1,040	17	—
Patwardhan and Dhawan (2004)	Nevoscope	580, 610	2	512 × 512
Patwardhan, Dhawan, and Relue (2005)	Nevoscope	580, 610	2	512 × 512
Tomatis et al. (2005)	SpectroShade	483–950	15	640 × 480
Carrara et al. (2007)	SpectroShade	483–950	15	640 × 480
Kazianka, Leitner, and Pilz (2008)	Custom	—	300	640 × 480
Świtoński, Michalak, Josiński, and Wojciechowski (2010)	VariSpec	410–710	21	—
Nagaoka, Nakamura, Kiyohara, and Sota (2012)	ImSpector V8E	380–780	124	512 × 512
Nagaoka, Nakamura, Okutani, Kiyohara, and Sota (2012)	ImSpector V8E	380–780	124	512 × 512
Suárez et al. (2012)	Custom	400–1,100	—	—
Nagaoka et al. (2013)	ImSpector V8E	380–780	124	512 × 512
Quinzán et al. (2013)	Custom	400–1,100	71	640 × 480
Nagaoka et al. (2015)	ImSpector V8E	450–750	124	1,024 × 768
Zheludev, Pölönen, Neittaanmäki-Perttu, and Averbuch (2015)	VTT/Revenio	500–885	76	320 × 240
Lorencs, Sinica-Sinavskis, Jakovels, and Mednieks (2016)	Nuance EX	450–950	51	—
Song et al. (2016)	MelaFind	430–950	10	1,280 × 1,024
Zherdeva et al. (2016)	STC UI RAS	450–750	61	1920 × 1,200
Stamnes et al. (2017)	Custom	365–1,000	10	—
Lihacova et al. (2018)	Custom	405–964	4	—
Rey-Barroso et al. (2018)	Custom	414–1,613	14	512 × 512

Publications denoted in bold indicate that the research is based on hyperspectral images. The "—" entries indicate that the information is not explicitly presented in the publication.

and SIAscope (Moncrieff, Cotton, Claridge, & Hall, 2002), both operating in the 400–1,000 nm spectral range. We are currently not aware of any commercially available hyperspectral imaging systems designed for skin lesion analysis.

Both multispectral and hyperspectral images are commonly represented as three-dimensional matrices (or data cubes), where the first two axes represent the spatial coordinates, and the third axis contains the spectral band measurements. There are two commonly used approaches to visualizing the information stored in a hyperspectral image. The first one is to pick one or more pixels (spatial coordinates) and plotting their respective spectral band measurements by wavelength. The other way is to visualize all pixels for one or more spectral bands as individual grayscale or color-mapped images. See Figure 5 for an example of both types of visualization.

There are several ways to capture both hyperspectral and multispectral images (Li et al., 2013), but from the perspective of data science applications, how images are captured is not crucial. However, what the captured image data represents is important. Both types of images contain information that represents either absorption, reflectance, or radiance at specific wavelengths across the electromagnetic spectrum. Measurements at discrete wavelengths are typically not performed, but

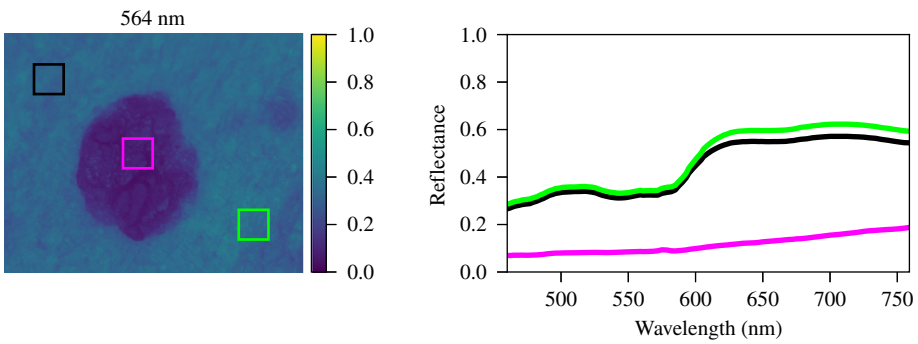


FIGURE 5 An example of what spectral curves for hyperspectral pixels can look like. The plot on the left shows a representation of a hyperspectral reflectance image at an arbitrarily chosen wavelength. On the right, the mean reflectance values are plotted, where the colors of the curves correspond to the colored regions in the reflectance image. The mean curves are calculated based on all pixels in each region

measurements are instead performed across narrow ranges of wavelengths referred to as spectral bands. Multispectral images are often captured at specifically chosen spectral bands across the supported spectral range of the camera. In many scenarios, the chosen spectral bands are picked based on prior knowledge, such as known absorption wavelengths of certain chemical compounds or similar. Other times the spectral bands are chosen somewhat arbitrarily, or at evenly spaced intervals across the entire spectral range. Commercial multispectral systems are typically capable of capturing 5–15 spectral bands across their supported spectral range. Because of spectral resolution and how wavelengths are typically chosen, captured multispectral data should be considered as consisting of discrete measurements. Hyperspectral images are captured with constant sampling rate across the spectral range of the camera, and can have hundreds of spectral bands depending upon the resolution of camera. Therefore, measurements in hyperspectral images are often considered to be continuous, which means that each pixel in a hyperspectral image can be said to represent a continuous spectral curve.

Before multispectral or hyperspectral images can be used as input to any classifier, statistical method, or other computational algorithm where images will be compared in some sense, they need to be preprocessed. One very important preprocessing step is calibration with respect to a known reference, typically an image of certified white reference material captured. The image of the white reference is captured right before or after taking an image of a skin lesion. This ensures that both images are captured under equivalent conditions. Certified white references used with hyperspectral systems have known spectral response, for example, 99.9% reflectance, across the entire supported spectral range, and are often intended to represent the maximum values measurable by a camera. In addition so-called dark current or dark reference images are usually also captured as part of the calibration process. These images can be captured by preventing light from hitting the camera sensor, and they therefore represent the minimum values measurable by a camera. An underlying assumption in this process is that the following inequality is fulfilled,

$$0 \leq I_{\text{dark}} < I_{\text{raw}} < I_{\text{white}}, \tag{1}$$

where I_{raw} is the raw image before calibration, I_{white} is the white reference image, and I_{dark} is the dark reference image.

A frequently used method for calibrating hyperspectral images is relative reflectance, which in this context is performed by rescaling spectral measurements from the skin lesion image with respect to the two reference images (Lawrence, Park, Windham, & Mao, 2003; W. Wang, Li, Tollner, Rains, & Gitaitis, 2012) captured under similar conditions. The relative reflectance image can be expressed as

$$I_{\text{reflectance}} = \frac{I_{\text{raw}} - I_{\text{dark}}}{I_{\text{white}} - I_{\text{dark}}}. \tag{2}$$

Given that the inequality in (1) is fulfilled, relative reflectance images will theoretically be bounded in (0, 1). This also implies that all calibrated images from the same camera system are comparable in a fairly robust sense since the process reduces the effects of the camera itself and the environment in which images are captured. Furthermore, images are scaled to the same reference domain.

Another calibration technique used in some skin lesion classification research is the so-called optical density (Lorencs et al., 2016; Zherdeva et al., 2016). In the context of multispectral and hyperspectral images, optical density can be defined as the logarithm of the ratio of a known reference image to the raw image,

$$I_{\text{OD}} = \log\left(\frac{I_{\text{reference}}}{I_{\text{raw}}}\right). \tag{3}$$

The reference image $I_{\text{reference}}$ can be a white reference image, or an image of some other reference material with known spectral characteristics.

In Rey-Barroso et al. (2018) a novel hyperspectral image calibration for skin analysis is presented. The first innovation is to employ a neutral-gray color of an X-Lite ColorChecker reference instead of a conventional certified white reference. The motivation is that this reference material exhibits reflectance characteristics closer to that of human skin across the spectral range. They also perform an additional calibration step designed to account for the influence of healthy skin, reportedly boosting the effects of malignant tissue.

4 | CLASSIFIER INPUT

The underlying goal of classification is to organize observations into two or more labeled classes. The classifier can be considered an algorithm that suggests a class affiliation based on the input characteristics of the observation. For early detection of skin cancer, there will typically only be two classes: malignant and benign. The classifier takes the skin lesion image, or features extracted from the image, as input and gives a binary output indicating whether the lesion is malignant or not.

The input to the classifier must contain information that makes it possible to discriminate according to the different classes. In the skin cancer situation, this means that the input must contain crucial properties of the skin lesion so that an image of a skin lesion can be assigned the correct class in a very robust manner. Since the input of the classifier plays such a crucial role, we will describe some important aspects of this for the skin cancer case.

4.1 | Feature extraction, feature selection, and dimensionality reduction

As pointed out earlier, a set of characteristics or features must be extracted from the image to construct a classifier. These features can be categorized into hand crafted features and summary statistical features. A third category, machine learned features, will not be discussed in this section since there are no deep learning classifiers yet for hyper- or multispectral skin lesions. However, we will discuss some aspects related to learning features from data in later sections.

The hand crafted features aspire at mimicking some aspect that is known to be discriminatory for lesion diagnosis, often inspired by, but not limited to, the ABCD rule of dermoscopy (Nachbar et al., 1994). Several hyper- and multispectral systems apply hand crafted features, exclusively or in combination with summary statistics features (Carrara et al., 2007; Stamnes et al., 2017; Tomatis et al., 2005).

The summary statistics features are typically the mean, variance, entropy, and so on, of the pixel value for each spectral band. Common for these features, and also some of the hand crafted features, is that the spatial information is not taken into account. Some systems use only the mean pixel value (Quinzán et al., 2013; Zherdeva et al., 2016), others use a different feature or a combination of summary statistics features (Lihacova et al., 2018; Lorencs et al., 2016; Nagaoka et al., 2015; Patwardhan et al., 2005; Rey-Barroso et al., 2018).

Each feature is calculated for each spectral band, and with a combination of a large set of features and many bands, dimensionality reduction can improve the performance of the system. With many spectral bands and/or features, some of the information is probably redundant, but each feature and band adds noise. Dimensionality reduction will reduce the noise and hence improve the classifier. If the number of images is small compared to the dimensionality of the images in a dataset, which is often the case for hyperspectral image datasets, a trained classifier will be unlikely to generalize well with regards to classifying samples not seen during training. The discriminatory power of a classifier initially increases as the number of feature dimensions increases, but then begins to decrease as the number of dimensions keeps increasing. This effect is often referred to as Hughes phenomenon (Shahshahani & Landgrebe, 1994). Therefore, dimensionality reduction is beneficial even if it reduces the amount of discriminatory information. The three main strategies for dimensionality reduction are band selection (Lorencs et al., 2016; Quinzán et al., 2013), feature subset selection (Patwardhan et al., 2005; Rey-Barroso et al., 2018; Stamnes et al., 2017), and principal component analysis (PCA) (Carrara et al., 2007; Kazianka et al., 2008; Tomatis et al., 2005). In spectral band selection and feature selection, a subset of the original bands and/or features is selected. This can be done by selecting a subset of bands, then the features are calculated for this subset (Lorencs et al., 2016; Quinzán et al., 2013). It can also be done in combination (Rey-Barroso et al., 2018), where the features are calculated for all bands and then the best band-feature pairs are selected, potentially keeping all spectral bands. The advantage of the first approach is that the number of bands are reduced, which can lead to a simpler camera construction in the future. In both approaches, the interpretability is kept intact. When PCA is employed to reduce the spatial or spectral dimensions, the result is a linear combination of features and spectral bands with different positive and negative weights, and the interpretability of the resulting PCA features is to some extent lost. It can be argued that interpretability is of lesser importance if the classifier is accurate enough, but so far there are no systems with accuracies high enough to justify a "black box" approach, given the potential fatal outcome of a misclassified melanoma.

4.2 | Selecting optimal spectral bands

A promising approach for dimensionality reduction of hyperspectral images is to reduce the number of spectral bands by selecting a subset of optimal wavelengths in a given hyperspectral image. This reduction can be performed by

focusing on the spectral dimension of the captured image. In hyperspectral imaging, two main approaches have been proposed to reduce dimensionality: selection of spectral features or selection of spatial features (Dai, Cheng, Sun, & Zeng, 2015). On one hand, spectral feature selection can be based on for example, correlation analysis of the spectral bands. Reduction of the feature set can then be achieved by selecting those bands that provide the most salient statistical information. Different search strategies have been proposed for spectral feature selection; complete, heuristic, or random search. These search strategies have been used in conjunction with many algorithms such as branch and bound (BB) (Nakariyakul & Casasent, 2007), PCA (Xing, Bravo, Jancsók, Ramon, & De Baerdemaeker, 2005), artificial neural networks (ElMasry, Wang, & Vigneault, 2009), and competitive adaptive reweighted sampling (CARS) (Wu & Sun, 2013). On the other hand, spatial feature selection is focused on the selection of relevant image characteristics (color, shape, etc.) to discriminate the spectral bands that contain most information about the desired features. Investigations focusing on spectral and spatial feature selection in hyperspectral images up until now have been very limited, likely due to the small number of studies that have been carried out in the field of skin cancer detection using hyperspectral images as a whole.

Practical implementation of a hyperspectral imaging system if often challenging due to the complexity and cost of a hyperspectral camera is capable of capturing several hundreds of bands. Recent publications have addressed new strategies for obtaining a feasible and practical technical solution by reducing the number of spectral bands or combining different finite spectral bands. A direct consequence of the reduction of the total amount of information processed is the reduction of the computational requirements in a given algorithm. Reducing the spectra can also enable algorithms to operate in near real time. The common approach up until now has been to obtain a hyperspectral image composed of hundreds of bands and then analyze which bands provides more information to classify and differentiate the skin tumor.

In Zherdeva et al.'s (2016) study, an experimental setup with hyperspectral images in the 450–750 nm range is employed to discriminate between skin cancers. Based on analysis of the collected images, it was determined that the most relevant differences between healthy tissue and skin cancer are located in the spectral bands 530–570 nm and 600–700 nm. These bands correspond to the absorption wavelengths of hemoglobin and melanin, respectively (Rubins, Zaharans, Ļihačova, & Spigulis, 2014). It has been reported that hemoglobin concentration and the ratio of melanin in skin lesion tissue can be important biological markers for melanoma detection (MacKinnon, Vasefi, & Farkas, 2014; Vasefi et al., 2016).

A melanoma discriminator based on few spectral channels is proposed in Lorencs et al. (2016). The spectral band selection principles are based on a correlation study of the information contained between pixel values in optical density images of each pair of bands. The triplet of spectral bands at 540, 640, and 740 nm and at 540, 640, and 840 nm were selected as they presented the highest correlation values.

The algorithms described represent promising approaches in achieving feasible technical implementations for dermoscopic systems. Spectral band reduction, without degrading the performance of classifying skin lesions, speeds up both training and inference of associated algorithms and this can lead to near real-time operation of the overall system. This is an important characteristic for practical applications in clinical settings.

It is worth mentioning that hyperspectral image feature selection applied to skin cancer detection has been used in very few studies. Therefore, future investigation must be carried out to demonstrate the conclusions reported in the initial studies. Furthermore, reported applications of PCA on hyperspectral skin lesion images have not been focused on selecting optimal spectral bands. By using PCA to reduce the spatial dimensionality, which means applying PCA on each individual spectral band of an image, it should be possible to study which spectral bands are most salient (Yamal et al., 2012).

5 | THE POWER OF SPATIOSPECTRAL INFORMATION

Treating individual hyperspectral pixels in an image as independent observations from the same patient has certain advantages and disadvantages, both in statistical methodologies and machine learning. An immediate advantage is that an approach where individual pixels are classified will yield much bigger datasets for both training and testing, even with quite few images if they have large spatial dimensions. As an example, a small dataset consisting of 10 multi- or hyperspectral images with $1{,}000 \times 1{,}000$ pixels, becomes a massive dataset of 10 million observations in a pixel-wise scheme. Using a pixel-wise approach to detect skin cancer has not been widely studied, but some research has been performed; skin lesion segmentation based on a pixel-wise scheme was done in Świtoński et al. (2010). One challenge of a pixel-wise scheme is how to balance the classes in the dataset, and how to ensure the subdivision into training and test sets are distributed in a representative way with respect to the original distribution, but still performed at random. Another challenge is getting accurate labels at the pixel level, which means that for each individual pixel a corresponding individual classification or diagnosis must be known.

Acquiring accurate, fine-grained labels at the pixel level is currently not feasible, and this means that training and testing supervised models with a pixel-wise scheme will be difficult. The lack of published research using pixel-wise approaches to skin cancer detection might be indicative of these challenges, suggesting that more research is needed in this area.

Although individual pixels can be treated as independent to some extent, in reality neighboring pixels in an image are spatially dependent. By not accounting for this, models and algorithms are deprived of salient information that could otherwise be used to improve their classification performance. As an illustrative example, a dermatologist applying the ABCD rule of dermoscopy will take into account all of the spatial information visible in a dermoscope or dermoscopic image. If only presented with individual pixels, without the spatial context, classifying the skin lesion would unquestionably be much more challenging. Therefore, statistical methods and machine learning models should also be trained with the same type of spatially dependent data, either as full images or image patches.

Just like recent successful machine learning algorithms designed for clinical and dermoscopic RGB, images of skin lesions are trained on full images with all three color channels (Esteva et al., 2017), exploiting the full potential of hyperspectral images which involves using all of the spatial and spectral information encapsulated within the images. While this has to some extent been done in other applications of hyperspectral imaging, such as remote sensing (Chen, Zhao, & Jia, 2015; Makantasis, Karantzalos, Doulamis, & Doulamis, 2015; Mughees, Ali, & Tao, 2017), we are not aware of any published research in the area of skin cancer detection where all of the spatiospectral information is used in a combined, fully contextual approach. The most common practice up until now has been to use hand-crafted features, summary statistics, and other lower-dimensional features. This can work reasonably well in some cases, but most, if not all, such approaches are incapable of fully accounting for the spatial and spectral context of detected patterns and features. Many deep learning models designed for image classification, for example, convolutional neural networks (CNN) (Krizhevsky, Sutskever, & Hinton, 2012), are specifically tailored to learn robust feature detectors from data. The learned feature detectors (sometimes referred to as feature maps) have important traits such as translation equivariance. In simplified terms, a feature detector that has learned to detect, for example, eyes, will give the same activation response regardless of the spatial location of the pixels comprising an eye, but if the pixels are spatially translated, the activation will be translated respectively. For example, two activations of an "eye" feature in an image is not enough to detect the presence of a face, but two such feature activations in close, spatial proximity is a much stronger indication of a face. This is essentially how most CNN-based models learn to detect objects by synthesizing feature detectors from one layer into increasingly complex features in the next layer (Zeiler & Fergus, 2014). In Esteva et al.'s (2017) study, they develop a deep learning model that detects and classifies skin lesions using clinical and dermoscopic RGB images. One key component of their work is employing a technique referred to as transfer learning (Pan & Yang, 2010). More specifically, they perform fine-tuning of a pretrained CNN model using a large dataset of RGB-based clinical and dermoscopic skin lesion images. Using this type of deep learning technique is feasible when the modality of the dataset used to train the original model is equivalent to the modality of the dataset used to perform the fine-tuning. No such pretrained models for hyperspectral images are publicly available. This is likely one of the primary reasons why there exists no published research using deep learning methods on hyperspectral images for skin cancer detection. A concrete example of the spatiospectral information richness and variation is depicted in Figure 6. From the figure it is clear that different physical properties of the lesion are captured at different wavelengths. How to exploit this information is not immediately obvious however. One suggestion is that such knowledge should be learned from data using deep learning models as opposed to being captured by hand-crafted feature extractors, or explicitly modeled in other ways.

Due to substantial differences in dataset modality, spatial dimensions, and number of channels/bands, transfer learning based on models trained on RGB images is not directly applicable to hyperspectral images. It has been shown feature detectors learned in the early layers of CNNs trained on RGB images are sensitive to colors (Zeiler & Fergus, 2014). In other words,

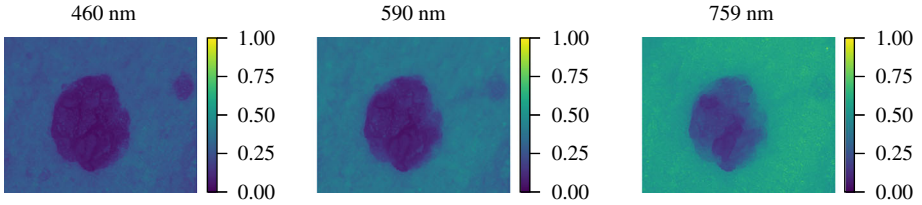


FIGURE 6 Examples of the different information captured in hyperspectral images at different wavelengths. Each image represents a reflectance image at a specific wavelength. Note how certain features appear and disappear at the various wavelengths. In particular, note how the small lesion visible near the right-most edge of the left image is almost invisible at higher wavelengths, and at higher wavelengths, smaller subregions and structures in the central lesion become visible

these feature detectors have adapted to specific characteristics of RGB images, which are not trivially transferable to hyperspectral images. Therefore, we believe that any model must either be trained from scratch, or novel RGB-to-hyperspectral transfer learning techniques must be developed. The number of trainable parameters in the most frequently used deep learning models designed for RGB images are in the order of 10^7, sometimes as high as 10^8. The number of trainable parameters in the first layer will increase substantially when the number of channels/bands in the images is increased by one or two orders of magnitude, which is the case when going from RGB to hyperspectral images. Additionally, many of the popular models are optimized for images with spatial dimensions around 250×250 pixels and three color channels. Hyperspectral images used for skin cancer detection have much higher spatial resolution and many more channels. Given these differences, model architectures should be augmented for high-resolution hyperspectral images. Examples of such design adjustments might include increasing the number of learned feature detectors in each layer of the model, and increasing the total number of layers. Increasing the complexity of the model translates into increasing the total number of trainable parameters.

Based on these observations, it is clear that training deep learning models for skin cancer detection using hyperspectral images from scratch will be challenging given current techniques and technology; it will require large amounts of computational power due to the high dimensionality of the images, and the high number of parameters being optimized during training. Training from scratch will also require sufficiently large datasets of high-quality, domain-specific, and representative observations in order for the model to generalize well at classification tasks. As far as we know, there are no publicly available datasets of hyperspectral skin lesion images that are sufficiently large to train deep learning models for skin lesion classification. The lack of publicly available datasets and pretrained models are likely the key challenges that explain the lack of published research on deep learning methods for skin cancer detection using hyperspectral images.

6 | CRITICAL REMARKS AND ANALYSIS OF PUBLISHED RESULTS

In the context of cancer detection, the ideal system provides a class label for each image in accordance with the actual pathology of the lesion in question. The gold standard for skin lesion diagnosis is histopathology for excised lesions and dermoscopic evaluation for nonexcised lesions. Note that a nonbiopsied lesion can only have a benign diagnosis, as suspicion of malignancy automatically leads to excision and histopathological examination. Due to the potential fatal consequences of misclassifying a melanoma as benign, even low level of suspicion leads to excision.

A system that aims at clinical relevance must either have melanoma sensitivity close to 100% combined with a reasonable specificity, or provide information that benefits the physician in the decision on whether to excise the lesion in question. Both objectives have shown to be hard to achieve, and so far no system can be said to have achieved either.

To predict the performance on future data, which do not have class labels, a system is tested on either an independent test set or by the use of cross validation. For the outcome to be valid, the test set must be independent of all aspects of the system development, from bandwidth selection to classifier parameter settings. In addition, the test set must be large enough for the result to be generalizable, and reflect the population from where the future data will be collected. These standards can be difficult to achieve due to the nature of the problem at hand: hyperspectral cameras are expensive, require training to operate, and melanomas are rare but fatal. This results in small datasets, and combined with high dimensionality there is often not enough data for sufficient training and adequate testing. The differences in imaging acquisition systems hinder combining different datasets.

Several publications report performance on the same set of data that were used to develop the system (Kazianka et al., 2008; Lihacova et al., 2018; Lorencs et al., 2016; Rey-Barroso et al., 2018; Zherdeva et al., 2016), which give highly optimistic results. This bias does not only apply when the test set is used to train the classification algorithm itself, but applies for all parts of system development, including bandwidth selection (Quinzán et al., 2013), feature selection (Patwardhan et al., 2005; Stamnes et al., 2017), and post hoc threshold settings for classification (Nagaoka et al., 2015; Patwardhan et al., 2005). The impact might not be obvious, but it is indisputable (Smialowski, Frishman, & Kramer, 2010).

In an effort to overcome the limitations of a small dataset, cross validation have been used (Nagaoka et al., 2015; Quinzán et al., 2013), but when using the entire dataset for bandwidth or feature selection, or parameter setting, the results are invalid.

It can be argued that incorrect use of statistical tools not necessarily disregards the results altogether, but the drop in performance is usually dramatic. The performance of MelaFind dropped from 85% specificity (Elbaum et al., 2001) to 9% specificity for near 100% sensitivity, when tested on a proper independent test set (Monheit et al., 2011). For more examples, see Møllersen et al. (2015).

The reported classification results with independent test sets are:

Publication	Sensitivity (%)	Specificity (%)	# Melanomas	# Lesions in total
Song et al. (2016)	50	23	4	55
Nagaoka et al. (2015)	75	97	24	132
Tomatis et al. (2005)	80	90	41	1,369
Carrara et al. (2007)	95	53	76	1,208

The study of Song et al. (2016) tested MelaFind in a clinical setting, but contains only four melanomas, and the results are therefore not generalizable. The Nagaoka et al. (2015) study consisted of 24 melanomas and 108 other skin lesions, but the lesions are from both patients and volunteers, and can therefore not be said to reflect any future population. Tomatis et al. (2005) had a large dataset with excised lesions consecutively collected, and in addition nonexcised lesions that were randomly collected in a clinical setting. Ideally, both the excised and nonexcised lesions should have been consecutively collected, but compared to other datasets in the field of computer-aided skin lesion classification, this dataset has high quality. The test set consisted of 41 melanomas and 306 nonmelanomas, confirmed by histopathology, and 1,022 lesions that were diagnosed as benign without excision. When using only the set of excised lesions, the specificity dropped to 77%, which clearly shows the enormous impact that the inclusion criteria for the dataset can have on the result. Carrara et al. (2007) reported in their study the sensitivity and specificity to whether a lesion should be excised, with the dermatologist's decision as ground truth. The numbers reported here are according to melanoma/nonmelanoma classes. Note the Tomatis et al. (2005) and Carrara et al. (2007) studies use overlapping datasets and methods.

The reported performance of a system will vary from one test set to another due to its random nature. The Clopper–Pearson confidence interval for the 95% sensitivity in the study of Carrara et al. (2007) is 87–99%, which clearly demonstrates the need for large test sets for reliable results.

The common practice of reporting of a single sensitivity-specificity pair makes comparison between systems impossible. The high specificity reported by Tomatis et al. (2005) drops when the sensitivity is increased, as shown in their receiver operating characteristic (ROC) curve, which shows the specificity as a function of sensitivity. The curve is not detailed enough to extract the exact numbers. The reported 80% sensitivity of Tomatis et al. (2005), which corresponds to missing one out of five melanomas, is not relevant for a system intended for clinical use. A 95% sensitivity corresponds to missing 1 out of 20 melanomas, and might still not be high enough. There is no consensus for a lower limit for acceptable melanoma sensitivity, and therefore, to make comparisons between systems possible, the range of corresponding specificities for sensitivities from 95% to 100% should be reported. As shown in Møllersen, Zortea, Schopf, Kirchesch, and Godtliebsen (2017), the criterion for comparing different systems has huge impact on the resulting ranking. Summary performance measures such as the area under the ROC curve (AUC), does not distinguish between the two types of misclassifications; a system can have high AUC even if its ability to detect skin cancer is poor. This is not suitable in settings where a false negative (misclassifying a melanoma as benign) has much graver consequences than a false positive (misclassifying a benign lesion as malignant).

7 | FUTURE RESEARCH DIRECTIONS

7.1 | Long-term goals

In recent years, early detection of skin cancer using RGB images has been research focus in a large number of publications, see for example, Oliveira, Papa, Pereira, and Tavares (2018). Although the findings presented in Codella et al. (2018) and Esteva et al. (2017) are very promising, they are still not able to outperform experienced dermatologists. Future research using RGB images will likely suffer from effects equivalent to the law of diminishing returns, and because of this additional information richness is crucial to boost classification results even further.

The ultimate goal is to obtain classification systems that can lower the number of deaths caused by skin cancer significantly. A successful classification system will benefit from research in the following two directions.

First, there is a need to acquire a large quantity of high-quality data for all relevant skin cancers to be able to develop a successful classification system. Any database for clinical evaluation should be large enough to be able to provide good generalization, and hence reflecting the high-variability of data. This generalization is even more challenging in skin analysis, where the interpatient variability across different pigmented skin lesions is also influenced by the different skin phenotypes. By acquiring RGB and hyperspectral images for all cases, it will also be possible to give a more objective answer to the proposed

importance of hyperspectral information. Clearly, it will take many years before a sufficient number of datasets are available, but with such datasets available, the Common Task Framework described by Donoho (2017) can be used to obtain the best possible classification systems. After such results are available, clinical testing needs to be carried out before the whole system can be put into use.

Second, patients can contribute to earlier detection of harmful skin lesions by keeping an eye on the evolution of their skin lesions. A natural first step is therefore to design a system that can be used for monitoring skin lesions. Ideally, such a system should be precise, affordable, easy to use and interpretable. By designing a system like this, early detection of skin cancer will hopefully be significantly improved since one of the reasons for skin cancer-related death is the lack of early treatment. A successful monitoring system may result in earlier and more effective treatment, thereby reducing the number of deaths.

7.2 | Short-term goals

Although there exist several papers (Qi, Xing, Foran, & Yang, 2011; Taghizadeh, Gowen, & O'Donnell, 2011) that indicate that hyperspectral images contain information beyond RGB images, it seems natural to start with careful analyses that show how much and in what way hyperspectral information contributes in various types of classification algorithms, based in both statistical methodologies and machine learning.

Spatial and hyperspectral information gives a natural link to spatiotemporal methods and it is therefore natural to look into how such methods can be useful in the present task. In particular, there are links to image sequences in other applications of medicine. One example is functional MRI where an important aim is to find areas of the brain connected to specific tasks. This may, for example, be crucial in connection with brain surgery. Similar ideas could potentially be used to find "suspicious areas" that may be an indicator of a serious change in a skin lesion. Research in this direction should be performed in close collaboration with dermatologists, and may turn out to be well worthwhile since it could give rise to a boost in early detection of skin cancer. Another possibility is to look for particular shapes or features in the hyperspectral curves, thereby giving rise to important new features in a future classification rule.

Clustering of the hyperspectral signatures that gives rise to specific RGB values will give a potential link between RGB and hyperspectral images. This will give important knowledge about how homogeneous such clusters are, and it may also lead to a better understanding of the extra information obtained by hyperspectral signatures.

When the research community has gathered a large number of images, these datasets may be used to learn the characteristics of each class. One important research area here would be to see if hyperspectral images could be used to better distinguish between melanoma and other types of skin cancer. This would be an extremely important result since melanomas are fatal, whereas some types of nonmelanoma skin lesions are considered harmless. Dermatologists are able to distinguish these classes well, but this can be a very difficult task for general practitioners.

Preliminary results (Li, Zhou, Liu, Wang, & Guo, 2015; Q. Wang, Wang, Zhou, Li, & Wang, 2017; Ortega et al., 2018) indicate that pathology results can be improved both with respect to precision and time using hyperspectral imaging. Further investigations are needed to confirm this and to get a better understanding of how this new technology can be beneficial for this purpose.

Analysis of dermatological hyperspectral images is in our opinion the most important area for research in the near future. Monitoring the evolution of skin lesions over time is an important part of such research. In addition, it is important to analyze hyperspectral images using a large number of statistical tools, thereby gaining more knowledge about such data and be in better position to design classification systems when sufficiently large datasets become available.

For future classification systems, finding optimal data representation is a key to success. Also known as feature learning, this is the task of finding a representation of the input that will result in the best possible performance of the classification algorithm (Bengio, Courville, & Vincent, 2013). In the application at hand, the skin lesion's state is partially represented by the curves measured by the hyperspectral camera. To this end, we seek a way to represent the rich data contained in the skin lesion state that will result in a successful algorithm.

Using deep learning for reducing the dimensionality of hyperspectral images is believed to be an important field of research. Instead of using methodologies based on variance analysis, entropy, or other information measures, we suggest that learning robust lower-dimensional representations of the data using for example, deep autoencoders (Hinton & Salakhutdinov, 2006) could lead to better classification performance. The spatiospectral information encoded in hyperspectral images is complex, and it is not immediately obvious that conventional methods such as PCA are sufficiently capable of capturing this. Furthermore, learning shared representations might make it feasible to combine hyperspectral skin lesion datasets (Ngiam et al.,

2011). Researching the potential gains of using deep learning approaches for dimensionality reduction could yield extremely important results.

The incorporation of scale-space ideas can also be explored in obtaining an efficient state space representation. Scale-space theory is a framework for representing signals on multiple scales, developed by the computer vision, image processing, and signal processing communities. Scale-space ideas could be used to select tuning parameters in the FDA approach. For a basis expansion representation of hyperspectral curves, for instance, several key parameters (e.g., bandwidth, degree of the derivative) must be selected (Chaudhuri & Marron, 2000). As the representation may be very sensitive to these parameters, scale-space methods can provide useful insight. For instance, SiZer is a visual tool to examine when the derivative of a scatterplot smoother is significantly negative, possibly zero, or significantly positive across a range of smoothing bandwidths (Chaudhuri & Marron, 1999).

8 | CONCLUDING REMARKS

Recent advances in hyperspectral imaging for skin cancer detection show great promise, and we believe that further research can lead to a significant reduction in the number of deaths caused by skin cancer. However, there are still many open research questions that must be addressed, such as what are the benefits of training classifiers with hyperspectral skin lesion images as opposed to clinical and dermoscopic images of skin lesions captured with conventional RGB cameras. To answer this, large, high-quality datasets of skin lesion images need to be collected using both hyperspectral and conventional RGB cameras. Importantly, both types of images need to be collected from all observed skin lesions in order to make it possible to perform, for example, statistical analysis, and to compare classification algorithms trained on both types of images. Once enough data has been collected, the data can be analyzed using statistical methodologies such as functional data analysis, multivariate analysis, and so on. Furthermore, classification algorithms can be trained using conventional statistical model-based methodologies and more recent developments based on deep learning approaches. How to architect and optimize algorithms and models for skin cancer detection using hyperspectral imaging need to be discovered by further research. Hyperspectral imaging is widely used in other fields of research such as remote sensing, and such research should provide a good foundation on which to build future research efforts toward skin cancer detection.

For reported performance results of classification systems to be valid and reliable, to ease comparison between systems, and to ensure that the clinical aspect is not ignored, we have the following recommendations for data collection and statistical analysis of the results:

1. Use an independent test set, not cross validation.
2. Report specificities for sensitivities from 95 to 100%.
3. Collect data in a clinical-like setting, with clearly stated inclusion and exclusion criteria. The data should be collected consecutively to reflect the underlying distribution of the population in question (e.g., hospital patients, primary care patients, etc.).
4. Report confidence intervals for the sensitivities.
5. If the available dataset is too small for independent test set, other aspects of the system such as spectral band selection or feature selection can be reported instead.

For a more detailed list that will increase the quality of a study even further, see Rosado et al. (2003).

ACKNOWLEDGMENTS

We would like to thank Dr. Herbert Kirchesch for his help in collecting all of the conventional RGB images and hyperspectral images used in our example figures. The camera prototype used for capturing the hyperspectral images used in our example figures was provided by Revenio Research Oy. This work has been supported in part by the Canary Islands Government through the ACIISI (Canarian Agency for Research, Innovation and the Information Society), ITHACA project "Hyperspectral Identification of Brain Tumors" under Grant Agreement ProID2017010164 and it has been partially supported also by the Spanish Government and European Union (FEDER funds) as part of support program in the context of Distributed HW/SW Platform for Intelligent Processing of Heterogeneous Sensor Data in Large Open Areas Surveillance Applications (PLATINO) project, under contract TEC2017-86722-C4-1-R. Additionally, this work was completed while Samuel Ortega was beneficiary of a predoctoral grant given by the "Agencia Canaria de Investigacion, Innovacion y Sociedad de la

Información (ACIISI)" of the "Conserjería de Economía, Industria, Comercio y Conocimiento" of the "Gobierno de Canarias", which is part financed by the European Social Fund (FSE) (POC 2014–2020, Eje 3 Tema Prioritario 74(85%)). Finally, this work has been also supported in part by the 2016 PhD Training Program for Research Staff of the University of Las Palmas de Gran Canaria. The project has also partially been funded by grant A33020 from Tromsø Forskningsstiftelse.

RELATED WIREs ARTICLES

Bayesian latent modeling of spatio-temporal variation in small-area health data
Modern perspectives on statistics for spatio-temporal data
Computational biology perspective: kernel methods and deep learning
Networks and pathways in pigmentation, health, and disease
Deep learning for remote sensing image classification: A survey

ORCID

Thomas Haugland Johansen https://orcid.org/0000-0003-3572-4706
Himar Fabelo https://orcid.org/0000-0002-9794-490X
Gustavo M. Callico https://orcid.org/0000-0002-3784-5504
Fred Godtliebsen https://orcid.org/0000-0001-7896-8634

REFERENCES

Abbasi, N. R., Shaw, H. M., Rigel, D. S., Friedman, R. J., McCarthy, W. H., Osman, I., … Polsky, D. (2004). Early diagnosis of cutaneous melanoma. *The Journal of the American Medical Association*, *292*(22), 2771–2776. https://doi.org/10.1001/jama.292.22.2771

Ahnlide, I., Bjellerup, M., Nilsson, F., & Nielsen, K. (2016). Validity of ABCD rule of dermoscopy in clinical practice. *Acta Dermato-Venereologica*, *96*(3), 367–372. https://doi.org/10.2340/00015555-2239

American Cancer Society. (2018). *Cancer facts and figures 2018.* Retrieved from https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2018.html

Annessi, G., Bono, R., Sampogna, F., Faraggiana, T., & Abeni, D. (2007). Sensitivity, specificity, and diagnostic accuracy of three dermoscopic algorithmic methods in the diagnosis of doubtful melanocytic lesions. *Journal of the American Academy of Dermatology*, *56*(5), 759–767. https://doi.org/10.1016/j.jaad.2007.01.014

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *00*(00), 1–31. https://doi.org/10.3322/caac.21492

Carrara, M., Bono, A., Bartoli, C., Colombo, A., Lualdi, M., Moglia, D., … Marchesini, R. (2007). Multispectral imaging and artificial neural network: Mimicking the management decision of the clinician facing pigmented skin lesions. *Physics in Medicine and Biology*, *52*(9), 2599–2613. https://doi.org/10.1088/0031-9155/52/9/018

Chaudhuri, P., & Marron, J. S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, *94*(447), 807–823. https://doi.org/10.1080/01621459.1999.10474186

Chaudhuri, P., & Marron, J. S. (2000). Scale space view of curve estimation. *The Annals of Statistics*, *28*(2), 408–428. https://doi.org/10.1214/aos/1016218224

Chen, Y., Zhao, X., & Jia, X. (2015). Spectral–spatial classification of hyperspectral data based on deep belief network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *8*(6), 2381–2392. https://doi.org/10.1109/JSTARS.2015.2388577

Codella, N. C. F., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., & Halpern, A. (2018). Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (pp. 168–172). Washington, DC: IEEE. https://doi.org/10.1109/ISBI.2018.8363547

Dai, Q., Cheng, J.-H., Sun, D.-W., & Zeng, X.-A. (2015). Advances in feature selection methods for hyperspectral image processing in food industry applications: A review. *Critical Reviews in Food Science and Nutrition*, *55*(10), 1368–1382. https://doi.org/10.1080/10408398.2013.871692

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745–766. https://doi.org/10.1080/10618600.2017.1384734

Elbaum, M., Kopf, A. W., Rabinovitz, H. S., Langley, R. G., Kamino, H., Mihm, M. C., … Wang, S. (2001). Automatic differentiation of melanoma from melanocytic nevi with multispectral digital dermoscopy: A feasibility study. *Journal of the American Academy of Dermatology*, *44*(2), 207–218. https://doi.org/10.1067/mjd.2001.110395

ElMasry, G., Wang, N., & Vigneault, C. (2009). Detecting chilling injury in red delicious apple using hyperspectral imaging and neural networks. *Postharvest Biology and Technology*, *52*(1), 1–8. https://doi.org/10.1016/j.postharvbio.2008.11.008

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056

Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., . . . Bray, F. (2013). *GLOBOCAN 2012 v1.0, cancer incidence and mortality worldwide: IARC cancerbase no. 11 [internet].* Technical report, international agency for research on cancer. Retrieved from http://globocan.iarc.fr

Hinton, G. E., & Salakhutdinov, R. R. (2006). *Reducing the dimensionality of data with neural networks.* Technical Report No. 5786. https://doi.org/10.1126/science.1127647

Jet Propulsion Laboratory, California Institute of Technology. *Airborne Visible InfraRed Imaging Spectrometer (AVIRIS)—Imaging Spectroscopy.* Retrieved from https://aviris.jpl.nasa.gov/aviris/imaging%7B%5C_%7Dspectroscopy.html

Kazianka, H., Leitner, R., & Pilz, J. (2008). Segmentation and classification of hyper-spectral skin data. *Data Analysis, Machine Learning and Applications*, *31*, 245–252. https://doi.org/10.1007/978-3-540-78246-9_29

Korotkov, K., & Garcia, R. (2012). Computerized analysis of pigmented skin lesions: A review. *Artificial Intelligence in Medicine*, *56*(2), 69–90. https://doi.org/10.1016/j.artmed.2012.08.002

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25 (nips 2012)*. https://doi.org/10.1145/3065386

Kupetsky, E. A., & Ferris, L. K. (2013). The diagnostic evaluation of MelaFind multi-spectral objective computer vision system. *Expert Opinion on Medical Diagnostics*, *7*(4), 405–411. https://doi.org/10.1517/17530059.2013.785520

Lachenal, G., & Ozaki, Y. (1999). Advantages of near infrared spectroscopy for the analysis of polymers and composites. *Macromolecular Symposia*, *141*(1), 283–292. https://doi.org/10.1002/masy.19991410123

Lawrence, K. C., Park, B., Windham, W. R., & Mao, C. (2003). Calibration of a pushbroom hyperspectral imaging system for agricultural inspection. *Transactions of the ASAE*, *46*(2), 513–521. https://doi.org/10.13031/2013.12940

Lee, K.-S., Cohen, W. B., Kennedy, R. E., Maiersperger, T. K., & Gower, S. T. (2004). Hyperspectral versus multispectral data for estimating leaf area index in four different biomes. *Remote Sensing of Environment*, *91*(3-4), 508–520. https://doi.org/10.1016/j.rse.2004.04.010

Li, Q., He, X., Wang, Y., Liu, H., Xu, D., & Guo, F. (2013). Review of spectral imaging technology in biomedical engineering: Achievements and challenges. *Journal of Biomedical Optics*, *18*(10), 100901. https://doi.org/10.1117/1.JBO.18.10.100901

Li, Q., Zhou, M., Liu, H., Wang, Y., & Guo, F. (2015). Red blood cell count automation using microscopic Hyperspectral imaging technology. *Applied Spectroscopy*, *69*(12), 1372–1380. https://doi.org/10.1366/14-07766

Lihacova, I., Bolochko, K., Plorina, E. V., Lange, M., Lihachev, A., Bliznuks, D., & Derjabo, A. (2018). A method for skin malformation classification by combining multispectral and skin autofluorescence imaging. In J. Popp, V. V. Tuchin, & F. S. Pavone (Eds.), *Biophotonics: Photonic solutions for better health care VI* (Vol. 1068535, p. 113). Strasbourg, France: SPIE. https://doi.org/10.1117/12.2306203

Lorencs, A., Sinica-Sinavskis, J., Jakovels, D., & Mednieks, I. (2016). Melanoma-nevus discrimination based on image statistics in few spectral channels. *Elektronika ir Elektrotechnika*, *22*(2), 66–72. https://doi.org/10.5755/j01.eie.22.2.12173

Lu, G., & Fei, B. (2014). Medical hyperspectral imaging: A review. *Journal of Biomedical Optics*, *19*(1), 010901. https://doi.org/10.1117/1.JBO.19.1.010901

MacKinnon, N., Vasefi, F., & Farkas, D. L. (2014). Toward in vivo diagnosis of skin cancer using multimode imaging dermoscopy: (I) clinical system development and validation. Paper presented at the *Proceedings of the SPIE 8947, Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues XII, 89470I*, (4 March 2014). https://doi.org/10.1117/12.2041818

Makantasis, K., Karantzalos, K., Doulamis, A., & Doulamis, N. (2015). Deep supervised learning for hyperspectral data classification through convolutional neural networks. In *2015 IEEE international geoscience and remote sensing symposium (IGARSS)*, (Vol. 2015-November, pp. 4959–4962). IEEE. https://doi.org/10.1109/IGARSS.2015.7326945

Møllersen, K., Kirchesch, H., Zortea, M., Schopf, T. R., Hindberg, K., & Godtliebsen, F. (2015). Computer-aided decision support for melanoma detection applied on melanocytic and nonmelanocytic skin lesions: A comparison of two systems based on automatic analysis of Dermoscopic images. *BioMed Research International*, *2015*, 1–8. https://doi.org/10.1155/2015/579282

Møllersen, K., Zortea, M., Schopf, T. R., Kirchesch, H., & Godtliebsen, F. (2017). Comparison of computer systems and ranking criteria for automatic melanoma detection in dermoscopic images. *PLoS One*, *12*(12), e0190112. https://doi.org/10.1371/journal.pone.0190112

Moncrieff, M., Cotton, S., Claridge, E., & Hall, P. (2002). Spectrophotometric intracutaneous analysis: A new technique for imaging pigmented skin lesions. *British Journal of Dermatology*, *146*(3), 448–457. https://doi.org/10.1046/j.1365-2133.2002.04569.x

Monheit, G., Cognetta, A. B., Ferris, L., Rabinovitz, H., Gross, K., Martini, M., … Gutkowicz-Krusin, D. (2011). The performance of MelaFind: A prospective multicenter study. *Archives of Dermatology*, *147*(2), 188–194. https://doi.org/10.1001/archdermatol.2010.302

Mughees, A., Ali, A., & Tao, L. (2017). Hyperspectral image classification via shape-adaptive deep learning. In *2017 IEEE international conference on image processing (ICIP)*, (Vol. 2017, September, pp. 375–379). IEEE. https://doi.org/10.1109/ICIP.2017.8296306

Nachbar, F., Stolz, W., Merkle, T., Cognetta, A. B., Vogt, T., Landthaler, M., … Plewig, G. (1994). The ABCD rule of dermatoscopy. {H}igh prospective value in the diagnosis of doubtful melanocytic skin lesions. *Journal of the American Academy of Dermatology*, *30*(4), 551–559. https://doi.org/10.1016/S0190-9622(94)70061-3

Nagaoka, T., Kiyohara, Y., Koga, H., Nakamura, A., Saida, T., & Sota, T. (2015). Modification of a melanoma discrimination index derived from hyperspectral data: A clinical trial conducted in 2 centers between March 2011 and December 2013. *Skin Research and Technology*, *21*(3), 278–283. https://doi.org/10.1111/srt.12188

Nagaoka, T., Nakamura, A., Kiyohara, Y., & Sota, T. (2012). Melanoma screening system using hyperspectral imager attached to imaging fiberscope. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, *30*, 3728–3731. https://doi.org/10.1109/EMBC.2012.6346777

Nagaoka, T., Nakamura, A., Okutani, H., Kiyohara, Y., Koga, H., Saida, T., & Sota, T. (2013). Hyperspectroscopic screening of melanoma on acral volar skin. *Skin Research and Technology*, *19*(1), 290–296. https://doi.org/10.1111/j.1600-0846.2012.00642.x

Nagaoka, T., Nakamura, A., Okutani, H., Kiyohara, Y., & Sota, T. (2012). A possible melanoma discrimination index based on hyperspectral data: A pilot study. *Skin Research and Technology*, *18*(3), 301–310. https://doi.org/10.1111/j.1600-0846.2011.00571.x

Nakariyakul, S., & Casasent, D. P. (2007). Adaptive branch and bound algorithm for selecting optimal features. *Pattern Recognition Letters*, *28*(12), 1415–1427. https://doi.org/10.1016/j.patrec.2007.02.015

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal Deep Learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)* (pp. 689–696) Bellevue, Washington, DC.

Oliveira, R. B., Papa, J. P., Pereira, A. S., & Tavares, J. M. R. S. (2018). Computational methods for pigmented skin lesion classification in images: Review and future trends. *Neural Computing and Applications*, *29*(3), 613–636. https://doi.org/10.1007/s00521-016-2482-6

Ortega, S., Fabelo, H., Camacho, R., de la Luz Plaza, M., Callico, G. M., & Sarmiento, R. (2018). Detecting brain tumor in pathological slides using hyperspectral imaging. *Biomedical Optics Express*, *9*(2), 818–831. https://doi.org/10.1364/BOE.9.000818

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Patwardhan, S. V., & Dhawan, A. P. (2004). Multi-spectral imaging and analysis for classification of melanoma. *Conference proceedings:... Annual International Conference of the* IEEE engineering in medicine and biology Society. *IEEE Engineering in Medicine and Biology Society*, *1*, 503–506. https://doi.org/10.1109/IEMBS.2004.1403204

Patwardhan, S. V., Dhawan, A. P., & Relue, P. A. (2005). Monte Carlo simulation of light-tissue interaction: Three-dimensional simulation for trans-illumination-based imaging of skin lesions. *IEEE Transactions on Biomedical Engineering*, *52*(7), 1227–1236. https://doi.org/10.1109/TBME.2005.847546

Qi, X., Xing, F., Foran, D. J., & Yang, L. (2011). Comparative performance analysis of stained histopathology specimens using RGB and multispectral imaging. Paper presented at the *Proceedings of the SPIE 7963, Medical Imaging 2011: Computer-Aided Diagnosis, 79633B*, (9 March 2011). https://doi.org/10.1117/12.878325

Quinzán, I., Sotoca, J. M., Latorre-Carmona, P., Pla, F., García-Sevilla, P., & Boldó, E. (2013). Band selection in spectral imaging for non-invasive melanoma diagnosis. *Biomedical Optics Express*, *4*(4), 514–519. https://doi.org/10.1364/BOE.4.000514

Rey-Barroso, L., Burgos-Fernández, F., Delpueyo, X., Ares, M., Royo, S., Malvehy, J., … Vilaseca, M. (2018). Visible and extended near-infrared multispectral imaging for skin cancer diagnosis. *Sensors*, *18*(5), 1441. https://doi.org/10.3390/s18051441

Rosado, B., Menzies, S., Harbauer, A., Pehamberger, H., Wolff, K., Binder, M., & Kittler, H. (2003). Accuracy of computer diagnosis of melanoma: A quantitative meta-analysis. *Archives of Dermatology*, *139*(3), 361–367. https://doi.org/10.1001/archderm.139.3.361

Rubins, U., Zaharans, J., Ļihačova, I., & Spigulis, J. (2014). Multispectral video-microscope modified for skin diagnostics. *Latvian Journal of Physics and Technical Sciences*, *51*(5), 65–70. https://doi.org/10.2478/lpts-2014-0031

Shahshahani, B. M., & Landgrebe, D. A. (1994). The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. *IEEE Transactions on Geoscience and Remote Sensing*, *32*(5), 1087–1095. https://doi.org/10.1109/36.312897

Smialowski, P., Frishman, D., & Kramer, S. (2010). Pitfalls of supervised feature selection. *Bioinformatics*, *26*(3), 440–443. https://doi.org/10.1093/bioinformatics/btp621

Smith, R. B. (2012). *Introduction to hyperspectral imaging.* Retrieved June 26, 2018, from https://www.microimages.com/documentation/Tutorials/hyprspec.pdf

Song, E., Grant-Kels, J. M., Swede, H., D'Antonio, J. L., Lachance, A., Dadras, S. S., … Rothe, M. J. (2016). Paired comparison of the sensitivity and specificity of multispectral digital skin lesion analysis and reflectance confocal microscopy in the detection of melanoma in vivo: A cross-sectional study. *Journal of the American Academy of Dermatology*, *75*(6), 1187–1192.e2. https://doi.org/10.1016/j.jaad.2016.07.022

Stamnes, J. J., Ryzhikov, G., Biryulina, M., Hamre, B., Zhao, L., & Stamnes, K. (2017). Optical detection and monitoring of pigmented skin lesions. *Biomedical Optics Express*, *8*(6), 2946–2964. https://doi.org/10.1364/BOE.8.002946

Suárez, I. Q., Carmona, P. L., García-Sevilla, P., Boldo, E., Pla, F., Jiménez, V. G., Lozoya, R., & de Lucía, G. P. (2012). Non-invasive Melanoma Diagnosis using Multispectral Imaging. In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, (January, pp. 386–393). SciTePress–Science. https://doi.org/10.5220/0003843803860393

Świtoński, A., Michalak, M., Josiński, H., & Wojciechowski, K. (2010). Detection of tumor tissue based on the multispectral imaging. In *International conference on computer vision and graphics* (Vol. 1732, pp. 325–333). https://doi.org/10.1007/978-3-642-15907-7_40

Taghizadeh, M., Gowen, A. A., & O'Donnell, C. P. (2011). Comparison of hyperspectral imaging with conventional RGB imaging for quality evaluation of Agaricus bisporus mushrooms. *Biosystems Engineering*, *108*(2), 191–194. https://doi.org/10.1016/j.biosystemseng.2010.10.005

Tomatis, S., Bono, A., Bartoli, C., Carrara, M., Lualdi, M., Tragni, G., & Marchesini, R. (2003). Automated melanoma detection: Multispectral imaging and neural network approach for classification. *Medical Physics*, *30*(2), 212–221. https://doi.org/10.1118/1.1538230

Tomatis, S., Carrara, M., Bono, A., Bartoli, C., Lualdi, M., Tragni, G., … Marchesini, R. (2005). Automated melanoma detection with a novel multispectral imaging system: Results of a prospective study. *Physics in Medicine and Biology*, *50*(8), 1675–1687. https://doi.org/10.1088/0031-9155/50/8/004

Tuia, D., Volpi, M., Copa, L., Kanevski, M., & Munoz-Mari, J. (2011). A survey of active learning algorithms for supervised remote sensing image classification. *IEEE Journal of Selected Topics in Signal Processing*, *5*(3), 606–617. https://doi.org/10.1109/JSTSP.2011.2139193

Unlu, E., Akay, B. N., & Erdem, C. (2014). Comparison of dermatoscopic diagnostic algorithms based on calculation: The ABCD rule of dermatoscopy, the seven-point checklist, the three-point checklist and the CASH algorithm in dermatoscopic evaluation of melanocytic lesions. *The Journal of Dermatology*, *41*(7), 598–603. https://doi.org/10.1111/1346-8138.12491

Vasefi, F., MacKinnon, N., Saager, R., Kelly, K. M., Maly, T., Booth, N., … Farkas, D. L. (2016). Separating melanin from hemodynamics in nevi using multimode hyperspectral dermoscopy and spatial frequency domain spectroscopy. *Journal of Biomedical Optics*, *21*(11), 114001. https://doi.org/10.1117/1.JBO.21.11.114001

Vestergaard, M. E., & Menzies, S. W. (2008). Automated diagnostic instruments for cutaneous melanoma. *Seminars in Cutaneous Medicine and Surgery*, *27*(1), 32–36.

Wang, Q., Wang, J., Zhou, M., Li, Q., & Wang, Y. (2017). Spectral-spatial feature-based neural network method for acute lymphoblastic leukemia cell identification via microscopic hyperspectral imaging technology. *Biomedical Optics Express*, *8*(6), 3017–3028. https://doi.org/10.1364/BOE.8.003017

Wang, W., Li, C., Tollner, E. W., Rains, G. C., & Gitaitis, R. D. (2012). A liquid crystal tunable filter based shortwave infrared spectral imaging system: Calibration and characterization. *Computers and Electronics in Agriculture*, *80*, 135–144. https://doi.org/10.1016/j.compag.2011.09.003

Wu, D., & Sun, D.-W. (2013). Application of visible and near infrared hyperspectral imaging for non-invasively measuring distribution of water-holding capacity in salmon flesh. *Talanta*, *116*, 266–276. https://doi.org/10.1016/j.talanta.2013.05.030

Xing, J., Bravo, C., Jancsók, P. T., Ramon, H., & De Baerdemaeker, J. (2005). Detecting bruises on 'golden delicious' apples using hyperspectral imaging with multiple wavebands. *Biosystems Engineering*, *90*(1), 27–36. https://doi.org/10.1016/j.biosystemseng.2004.08.002

Yamal, J.-M., Zewdie, G. A., Cox, D. D., Neely Atkinson, E., Cantor, S. B., MacAulay, C., … Follen, M. (2012). Accuracy of optical spectroscopy for the detection of cervical intraepithelial neoplasia without colposcopic tissue information; a step toward automation for low resource settings. *Journal of Biomedical Optics*, *17*(4), 047002. https://doi.org/10.1117/1.JBO.17.4.047002

Zeiler, M. D., & Fergus, R. (2014). *Visualizing and understanding convolutional networks* (pp. 818–833). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-10590-1_53

Zheludev, V., Pölönen, I., Neittaanmäki-Perttu, N., & Averbuch, A. (2015). Biomedical signal processing and control delineation of malignant skin tumors by hyperspectral imaging using diffusion maps dimensionality reduction. *Biomedical Signal Processing and Control*, *16*, 48–60. https://doi.org/10.1016/j.bspc.2014.10.010

Zherdeva, L. A., Bratchenko, I. A., Myakinin, O. O., Moryatov, A. A., Kozlov, S. V., & Zakharov, V. P. (2016). In vivo hyperspectral imaging and differentiation of skin cancer. *10024*:100244G. https://doi.org/10.1117/12.2246433

How to cite this article: Johansen TH, Møllersen K, Ortega S, et al. Recent advances in hyperspectral imaging for melanoma detection. *WIREs Comput Stat*. 2020;12:e1465. https://doi.org/10.1002/wics.1465

Chapter 10.

Paper II

Early Detection of Change by Applying Scale-Space Methodology to Hyperspectral Images

Stig Uteng, Thomas Haugland Johansen, Jose Ignacio Zaballos, Samuel Ortega, Lasse Holmström, Gustavo M. Callico, Himar Fabello, Fred Godtliebsen

Published

applied sciences

MDPI

Article

Early Detection of Change by Applying Scale-Space Methodology to Hyperspectral Images

Stig Uteng [1,*,†], Thomas Haugland Johansen [2,†], Jose Ignacio Zaballos [3,†], Samuel Ortega [3,†], Lasse Holmström [4,†], Gustavo M. Callico [3,†], Himar Fabelo [3,†] and Fred Godtliebsen [2,†]

1 Department of Education and Pedagogy, UiT The Arctic University of Norway, 9019 Tromsø, Norway
2 Department of Mathematics and Statistics, UiT The Arctic University of Norway, 9019 Tromsø, Norway; thomas.h.johansen@uit.no (T.H.J.); fred.godtliebsen@uit.no (F.G.)
3 Institute for Applied Microelectronics (IUMA), University of Las Palmas de Gran Canaria (ULPGC), 35001 Las Palmas de Gran Canaria, Spain; josegt@gmail.com (J.I.Z.); sortega@iuma.ulpgc.es (S.O.); gustavo@iuma.ulpgc.es (G.M.C.); hfabelo@iuma.ulpgc.es (H.F.)
4 Research Unit of Mathematical Sciences, University of Oulu, 90570 Oulu, Finland; lasse.holmstrom@oulu.fi
* Correspondence: stig.uteng@uit.no
† These authors contributed equally to this work.

Received: 30 December 2019; Accepted: 23 March 2020; Published: 27 March 2020

Abstract: Given an object of interest that evolves in time, one often wants to detect possible changes in its properties. The first changes may be small and occur in different scales and it may be crucial to detect them as early as possible. Examples include identification of potentially malignant changes in skin moles or the gradual onset of food quality deterioration. Statistical scale-space methodologies can be very useful in such situations since exploring the measurements in multiple resolutions can help identify even subtle changes. We extend a recently proposed scale-space methodology to a technique that successfully detects such small changes and at the same time keeps false alarms at a very low level. The potential of the novel methodology is first demonstrated with hyperspectral skin mole data artificially distorted to include a very small change. Our real data application considers hyperspectral images used for food quality detection. In these experiments the performance of the proposed method is either superior or on par with a standard approach such as principal component analysis.

Keywords: change detection; scale-space methodology; hyperspectral imaging

1. Introduction

For a time-varying system, detection of unexpected or unwanted change in its evolution can be of paramount importance. Examples include environmental monitoring, process control, or, referring to the examples considered in this article, identification of potentially malignant changes in skin moles or the onset of food quality deterioration (see, for example, [1–4]). The first changes may be small and manifest themselves in different scales and it may be crucial to detect them as early as possible. Statistical scale-space methodologies (see Section 2) can be very useful in such situations since exploring the measurements in multiple resolutions can help identify subtle changes. Examples of scale-space methods designed for change detection are the SiNos technique for capturing non-stationarities in a time series [5] and the iBSiZer method for detecting changes in images [6]. Our goal was to develop a method that can detect minor change while at the same time keeping the number of false alarms to a minimum. This is important in practical applications as a successful method must have both high sensitivity and high specificity.

Recently, Hindberg et al. proposed a scale-space method for testing whether k multivariate data sets of same dimension originate from the same distribution [7]. Thus, the proposed method solves the

classical *k*-sample problem using scale-space analysis and the method has proven successful in many applications. In the applications considered here the observed data consist of multivariate vectors obtained from spectral signatures and therefore changes in their characteristics can also be analyzed with this method. Unfortunately, it turns out that in this context the method suffers from two serious shortcomings: failing to detect very small changes and producing unacceptably high rates of false alarms in some situations (see Section 4). Our goal therefore is to design a scale-space method that would suffer less from these shortcomings.

As an illustration of the difficulty of detecting very small changes, consider the example in Figure 1 which is discussed in more detail in Sections 2 and 4. The original data set consists of a number of spectral signatures acquired by a push-broom hyperspectral camera, each signature corresponding to a particular spot in a skin mole. Several acquisitions of the mole are taken at the same time, and an example of one acquisition is given in Figure 1a where each curve corresponds to a specific spectral signature. To simulate a situation where the mole might begin to turn malignant, we manually distorted just one spectral signature (thus corresponding to a very small local change in the mole) in another acquisition of the same mole at spectral channel 80 on the horizontal axis in Figure 1a. In case of real moles, the first changes may be extremely hard to detect and a method with high sensitivity and specificity is therefore crucial. In our test, the distorted set of signatures in Figure 1b was compared with 14 other acquisitions and the goal was to detect the small change we manually introduced. It turned out that such a small change is indeed detected by our new methodology but not by the method suggested in [7] nor by a standard approach such as principal components analysis (PCA). We will return in more detail to this example in Section 4.

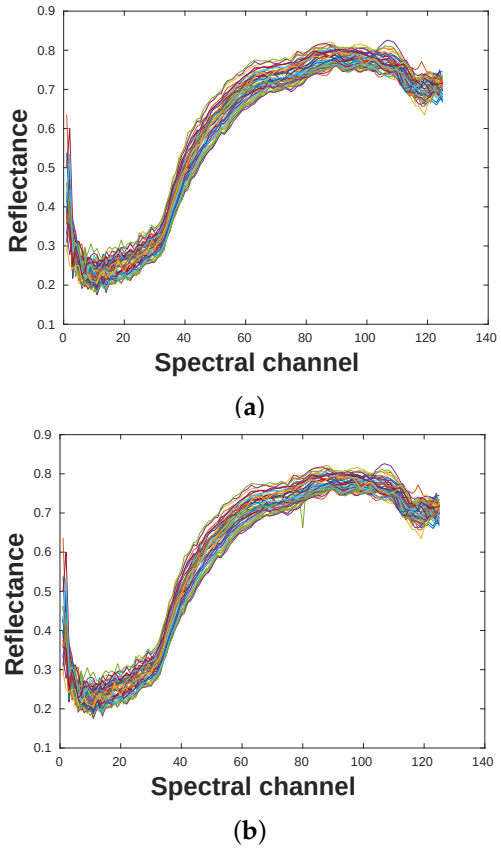


Figure 1. (**a**) The original undistorted curve families for the artificial example. (**b**) An example where a small artificial change has been introduced to the data set in Figure 1a at spectral channel 80.

2. Scale-Space Methodology

Scale-space theory is a framework for representing signals on multiple scales, developed by the computer vision, image processing and signal processing communities [8]. A recent review of statistical scale-space methodology can be found in [9]. The goal of statistical scale-space methodology is to extract statistically significant features from noisy data at several scales, often corresponding to different levels of resolution in the underlying object of interest. The data could be a set of observed curves where features at different levels of resolution might be of interest. These curves could, for example, correspond to spectral signatures from fish being frozen for different numbers of days, as is the case in our real data application. One acquisition of data consists of a number of p-dimensional vectors with unknown distribution, each vector representing the spectral signature at a particular pixel in the hyperspectral image. Thus, in our application, p represents the number of frequency bands (spectral channels) in the spectral signatures. In Section 4.2 we analyze three different acquisitions from the frozen fish. Under the null hypothesis, the number of days is assumed the same and the distributions are therefore assumed identical. In our approach, we perform several tests to flag when a new acquisition differs significantly from several previous acquisitions of day 0. The outcome of the tests is presented as a scale-space map, described in more detail below.

The core method of this paper is to test simultaneously for many different scales and positions (frequency bands). The scale s equals the number of different frequency bands being summed across. To be specific, this means that scale $s = 1$ corresponds to the situation where we test if the observed values at spectral frequency d are different between acquisitions of spectral signatures. At scale $s = 3$ and position d, a smoothing in terms of a weighted average of the observed values for spectral frequencies $d - 1$, d and $d + 1$ are used to test whether the acquisitions are different. The weights are calculated from an Epanechnikov kernel function (i.e., parabolic function) [10], the same as in [7]. For other scales, completely analogous smoothing over the frequency bands are made and used to perform the tests. Note that by applying this smoothing, we are able to test for differences in the acquisitions at all locations for a large number of scales. In fact, the tests are performed at all p spectral frequencies for a total number of n_s different scales. Instead of looking at a single location or a single scale, the described scale-space approach can help detect changes that appear at several levels of smoothing, i.e., resolution.

However, when testing for differences between spectral signature curves in different acquisitions it can be difficult to select the critical rejection thresholds due to multiple testing. One possibility is to use the Bonferroni correction method [11] designed to reduce false positives in testing multiple hypotheses. As an alternative, we also tried the statistical inference method described in [12] to find suitable critical rejection thresholds for the scale-space map. The critical values are used to test if a new acquisition differs from the existing acquisitions.

The training procedure at a location (d, s) is accomplished by comparing one acquisition to the others. To simplify the description we illustrate the methodology by testing for change in the sample mean, $\bar{X}$, over the pixels in the image. This training-procedure is the core difference between the method presented here and in [7], where there is a more direct comparison between curve families. Also, instead of the non-parametric Andersson-Darling test combined with either Bonferroni or False Discovery Rate correction for multiple hypothesis testing employed in [7], our novel method uses the t-test either with a Bonferroni correction or the inference approach suggested in [12]. Here, further, we assume that

$$\bar{X}_1 \sim N(\mu_1, \sigma^2) \tag{1}$$

for acquisition one and

$$\bar{X} = \frac{1}{n-1} \sum_{k=2}^{n} \bar{X}_k \sim N\left(\mu, \frac{\sigma^2}{n-1}\right), \tag{2}$$

for the remaining $n - 1$ of the acquisitions. Here n is the total number of acquisitions. The normal assumption makes sense due to the central limit theorem since all $\bar{X}_k$'s, $k = 1, \ldots, n$ are averages over a large number of observations. Note that this means that we perform the training procedure by leaving

one out cross validation. In the description above, the mean is chosen as parameter, but we have also implemented and performed our testing procedure for the median, the standard deviation and the range. We do this since these parameters can better describe certain aspects of a distribution and may therefore capture different types of changes. In practice, we will therefore typically test all these parameters for potential changes. For parameters other than the mean, Equation (1) will be an approximation that may be violated in practice. Equation (2) will, however, still be a reasonable approximation for all parameters due to the central limit theorem, but will sometimes only hold approximately.

In our description below, we estimate σ^2 by the standard estimator for variance using all acquisitions apart from the one left out. In the case acquisition 1 is left out, this means that σ^2 is estimated by

$$S^2 = \frac{1}{n-2}\sum_{i=2}^{n}(\bar{X}_i - \bar{X})^2.$$

The critical quantile at location (d,s), when only using the Bonferroni correction, is then given by

$$c(d,s) = t_{\frac{\alpha}{2p},n-2}, \tag{3}$$

where $\alpha = 0.05$ is the significance-level and p the number of spectral channels of each spectral signature curve. In addition to the Bonferroni-corrected quantile in Equation (3), we tried here the so-called global quantile

$$c(d,s)_G = \Phi^{-1}\left(\left(1-\frac{\alpha}{2}\right)^{\frac{1}{p\sum_{k=1}^{n_s}\theta_k}}\right), \tag{4}$$

proposed in [12]. Here Φ is the normal cumulative distribution function and n_s denotes the number of rows in the scale-space map. Moreover, θ_k is given by

$$\theta_k = 2\Phi\left(\frac{\sqrt{3\log p}}{2s_k}\right) - 1,$$

where s_k is the scale in row k. In the testing procedure, we test

$$H_0 : \mu_1 = \mu \text{ against } H_1 : \mu_1 \neq \mu$$

using the test statistic

$$T = \frac{\bar{X}_{\text{test}} - \bar{X}_{\text{train}}}{S\sqrt{1+\frac{1}{n-1}}},$$

where H_0 is rejected if

$$|T| > c(d,s) \text{ or } |T| > c(d,s)_G.$$

The algorithm is summarized in Algorithm 1 where *Par* is used to denote the parameter we are using in the tests.

The outcomes of the tests are graphically summarized in a scale-space map, where the horizontal and vertical axes correspond to spectral frequency bands and scales, respectively. Thus, at each location (d,s) we perform a test and the outcome is shown as a colored pixel, with red (blue) indicating a significant (not significant) difference at the position d for scale s.

To illustrate the method, consider the example introduced in Figure 1. Figure 2 shows the scale-space map produced by the procedure described above. The parameter used in this analysis was the range as it best detected the small change manually introduced to the data. Note how the map indicates a significant feature only for the smallest scales around the spectral channel given at point 80 on the horizontal axis. This is expected since the change is small and only present at one particular spectral channel for a single signature.

Algorithm 1 The SS_CC algorithm:

Initialization: Acquisitions that are correct under null hypothesis and the test-acquisitions are loaded.

procedure SS_CC_TRAIN()

Input: The loaded acquisitions that are correct under the null hypothesis.

Initialization: The significance level α is chosen.

for $i = 1 : n$ **do**

procedure LEAVE ONE OUT(k)

return index vector **v** without k

$\widehat{Par}(X_k) \sim N\left(Par, \sigma^2\right)$ from each (d,s) location.

for j in **v** **do**

$\widehat{Par}(X) \sim N\left(Par, \frac{\sigma^2}{n-1}\right)$ from each (d,s) location.

return $mean(\widehat{Par}(X))$, S

procedure SS_CC_TEST()

Input: The new acquisitions. $c(d,s)$, $c(d,s)_G$, $mean(\widehat{Par}(X))$, S

Initialization: The significance level α is chosen.

$T \leftarrow \frac{X_{test} - mean(\widehat{Par}(X))}{S\sqrt{1+\frac{1}{n-1}}}$

return Significance matrix for scale-space map

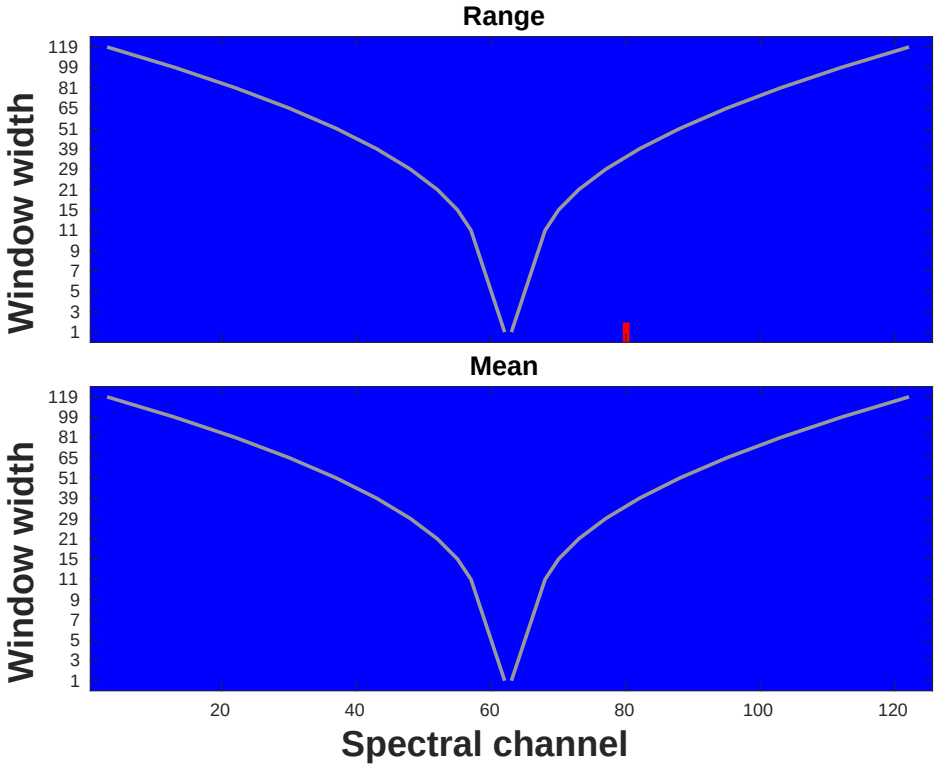


Figure 2. Scale-space significance map for the comparison between the hyperspectral image of a skin mole and an image obtained by manually distorting it. The original and distorted spectral signatures are shown in Figure 1. For the tests, the Hannig-Marron global rejection threshold was used both for the range and the mean.

3. Hyperspectral Acquisition System

In order to capture spectral signature curves from fish, a customized hyperspectral imaging (HSI) acquisition system was employed. Image acquisition was performed with a push-broom hyperspectral

camera with a spectral range of 410–1000 nm (see, for example, [3]) and spatial resolution of 0.3 mm across-track by 0.6 mm along-track (Norsk Elektro Optikk, model VNIR-1024). The camera was fitted with a lens focused at 1000 mm, mounted 1020 mm above a conveyor belt. Samples were illuminated using two custom made fiber optic line lights (Fiberoptics Technology inc., Pomfret, CT, USA), fitted with custom made collimating lenses yielding light lines approximately 5 mm wide (Optec S.P.A., Milano, Italy). Each line light was 400 mm wide, with six bundles of optical fibers. The light from 12 focused 150 W halogen lamps with aluminium reflectors (International Light Technologies, Peabody, MA, USA, model L1090) was fed into the fiberoptic bundles. The imaging and illumination setup is seen in Figure 3a. The optical power actually hitting the sample is approximately 0.16–0.79 Watt/(nm·sr·m^2).

The illumination system is composed of a controller unit which allows controlling the brightness and the light source. This system permits us to regulate the light intensity according to the sample characteristics, such as color, size or other parameters dependent on light. The acquisition technique employed by this camera is the so-called push-broom method, which consists of an optical system capturing an image from a line in a plane as depicted in Figure 3b. The camera collects images as seen in Figure 3b.

To capture a hyperspectral image, either the camera or the sample must be moved synchronously with the shoot of the camera. In this case, the sample is moved using a linear actuator by a stepper motor along a line. The light used has been tested to emit in the whole spectral range. Before starting the capturing process, the camera must be focused and calibrated with a dark reference and a white reference. In this process, a tile with 99% of reflectance was chosen for the white reference.

The spectral signatures of the same frozen fish are taken on day 0, day 2, day 4, day 7 and day 10. On each day, we captured 912,082 signature curves in each of the four acquisitions made. The four acquisitions from day 0 were then compared to the other acquisitions in order to find significant differences as described in Section 4.

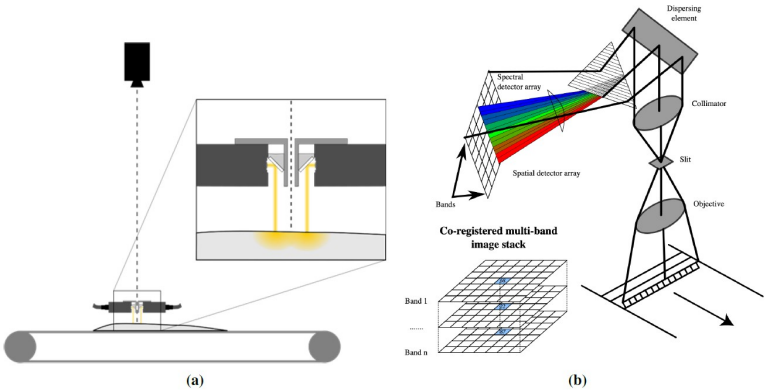


Figure 3. (**a**) The HSI setup with the Hyperspectral camera. © NOFIMA, Norway. (**b**) The HSI linear push-broom array. © https://commons.wikimedia.org/wiki/User:Arbeck.

After image acquisition, the data from the reference images were used to perform a radiometric calibration of the raw spectral signature of each pixel of the HSI cube as suggested in [13].

$$CI = \frac{RI - DI}{WI - DI}, \tag{5}$$

where CI is the calibrated image, RI is the raw image and WI and DI are the white and dark reference images, respectively.

4. Results

4.1. Artificial Mole Example

Recall from Section 1 that the data in our artificial data example were first obtained by acquiring a hyperspectral image of a skin mole and then modifying it manually in order to introduce a small

distortion that could simulate a change in the mole itself. The HSI system used for the mole example differs from the one described in Section 3 and a detailed description can be found in [14]. In our analysis, we compared the proposed novel technique to the method of Hindberg et al. described in [7]. When applied in the present context, this method uses a two-sample test to decide if the test sample distribution differs from the distribution under the null hypothesis. We also experimented with principal component analysis (PCA, e.g., [15]). By examining the scatter plots of the most important principal component directions we concluded that PCA is unable to detect the distortion in the data as seen in Figure 4.

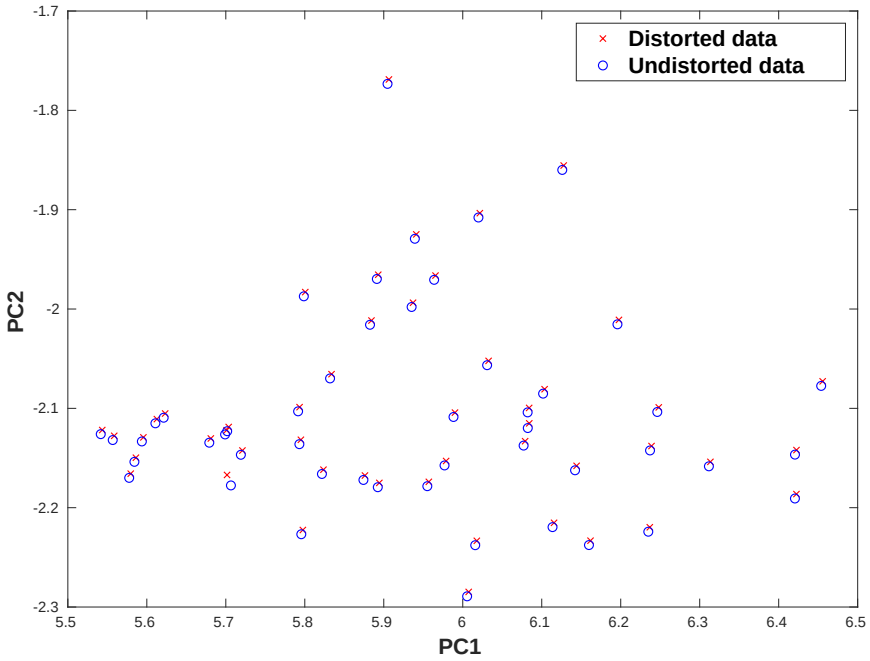
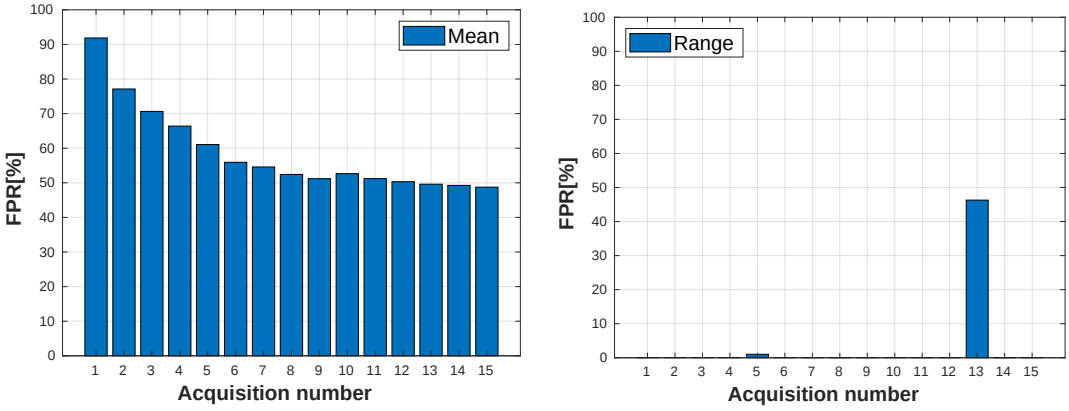


Figure 4. Plot of the first two principal components (PC) of the original mole spectral signatures (blue) and the spectral signatures of the distorted data (red) (see Figure 1b). Note the complete overlap of the two data sets, save for a small change at around (PC1, PC2) = (5.7, −2.2).

Because of this, we focus on a comparison between the new methodology and the one described in [7]. In general, the method in [7] performed poorly in this challenging situation. This was also reflected through the false positive rate (FPR) reported in Figure 5.



(**a**) (**b**)

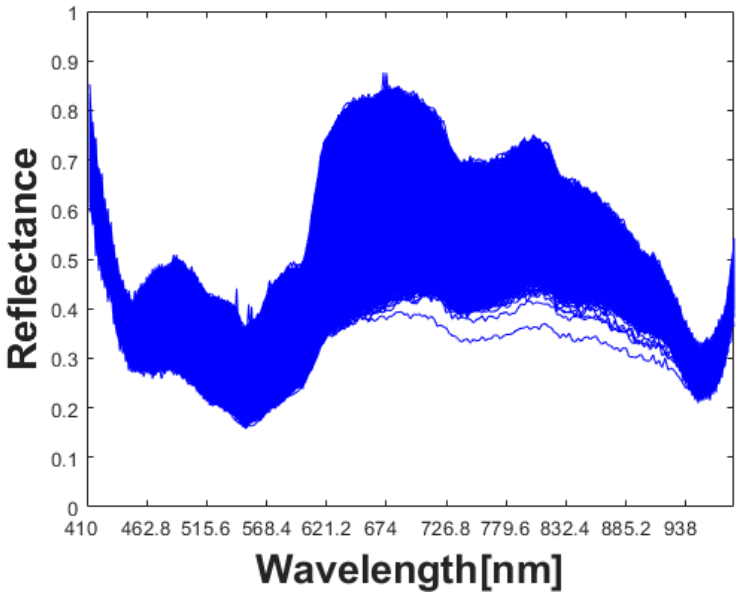
Figure 5. (**a**) The false positive rate (FPR), performing a leave-one-out test and using Bonferroni correction to account for multiple testing with the method described in [7]. (**b**) FPR obtained, performing a leave-one-out test and the Hannig-Marron global rejection threshold with the novel method proposed in this article.

The FPR results obtained by the scale-space methodology for the range are quite impressive. However, it should be noted that we were not able to get similar results for the mean, the median or the standard deviation. This is to be expected because the range is clearly best suited for detecting the type of distortion we introduced to the mole data reported in Figure 1. Still, while the range is the natural parameter to use in this situation, it is also clear that range based inference can be very sensitive to outliers, should the data include them.

4.2. Freshness of Fish

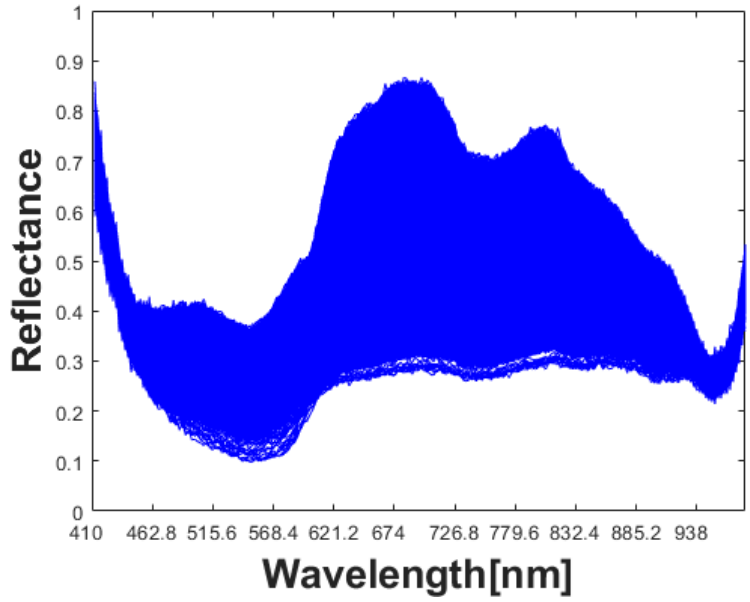
As a real data example, hyperspectral signatures of frozen fish were analyzed. The data acquisitions took place on several different days after a fish was captured. In our analysis, a subset of $301^2 = 90{,}601$ signatures were analyzed in each acquisition. This is a subset of the full HSI cube that consists of 912,082 spectral signature curves, chosen due to the upper limit of 130 GB of RAM available in the computer used to perform our experiments. An example taken from one fish at day 0 is given in Figure 6a and signatures for two different acquisitions for the same fish at day 4 are given in Figure 6b,c, respectively. Comparison of the signatures of Figure 6a with Figure 6b,c results in the significance maps in Figure 7. There the four panels show the results using two different parameters, the mean and the median. Changes are detected with the mean and the median, but with the standard deviation and the range, no statistically significant changes were detected. A careful examination of the curves in Figure 6 reveals that location parameters are expected to detect changes best in this case and this is indeed what happens. Typical FPR results are reported in Figure 8. From the results here we see that the Hannig-Marron critical value gives better performance. The difference is, however, not very clear in this example.

Due to computational challenges, results for the method of Hindberg et al. in [7] and PCA could not be obtained for the full data sets. In comparisons using only smaller subsets of the data, all three methods performed similarly in detecting changes while the their FPRs were similar to Figure 8.

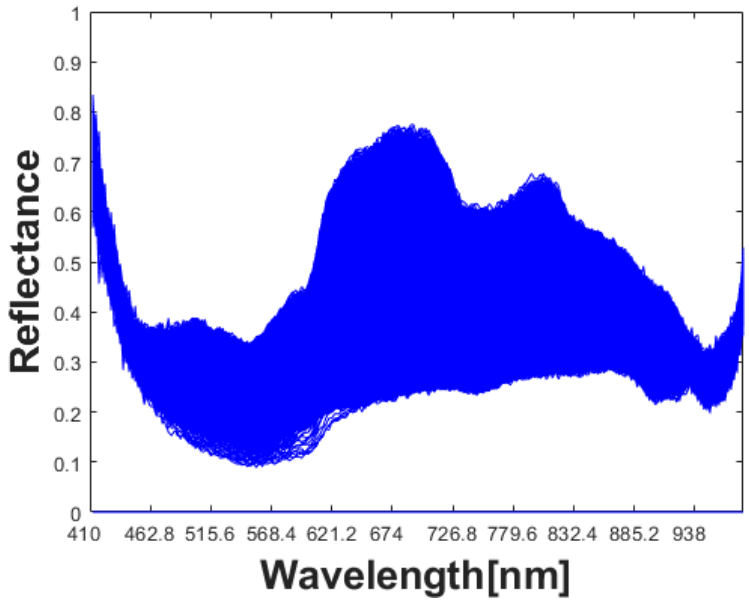


(**a**)

Figure 6. *Cont.*

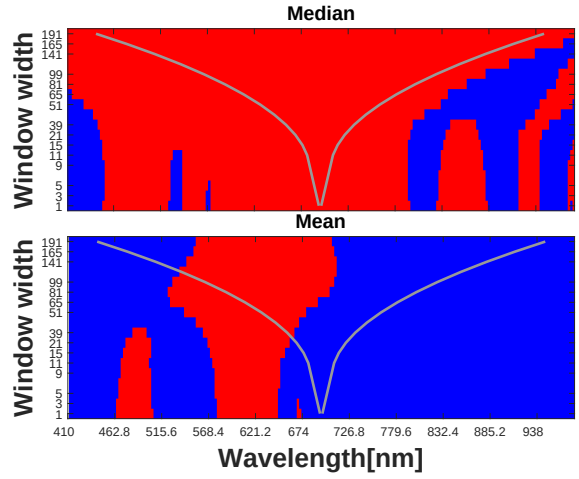


(**b**)

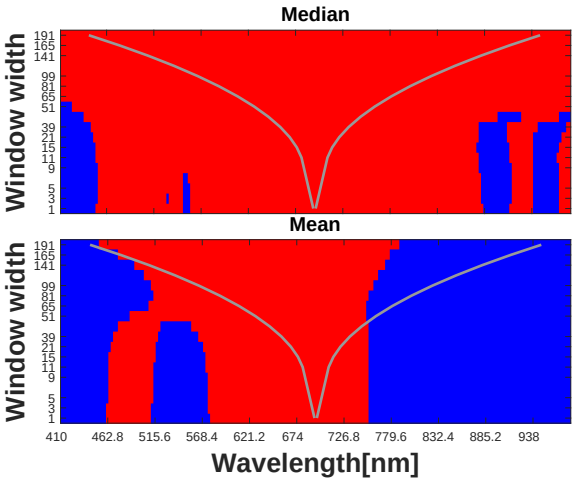


(**c**)

Figure 6. (**a**) Plot of HSI-curves from acquisition number one from frozen fish at day 0. (**b**) Plot of HSI-curves from acquisition number two from frozen fish at day 4. (**c**) Plot of HSI-curves from acquisition number four from frozen fish at day 4.

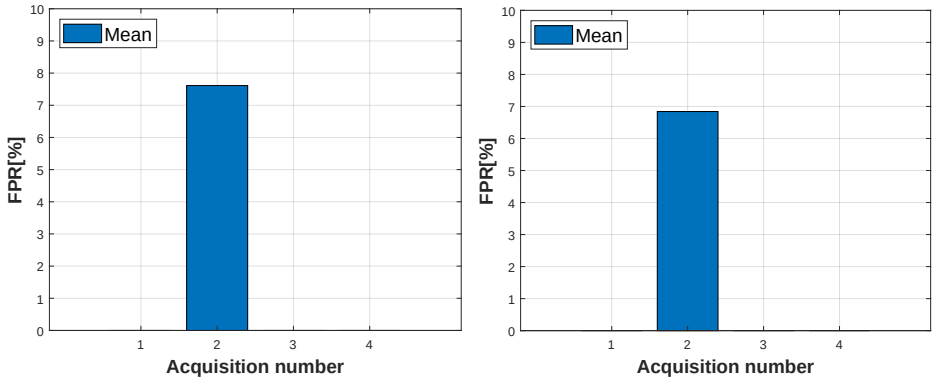


(**a**)



(**b**)

Figure 7. (**a**) Significance map for the comparison of the day 0 frozen fish acquisitions with day 4 acquisition number two using the median and the mean. (**b**) Significance map for the comparison of the day 0 frozen fish acquisitions with day 4 acquisition number four using the median and the mean.The Hannig-Marron rejection threshold was used in all maps.



(**a**) (**b**)

Figure 8. False positive rate (FPR) in the fish freshness example. (**a**) FPR from a leave-one-out test using Bonferroni correction to account for multiple testing. (**b**) FPR from a leave-one-out test and using the Hannig-Marron global rejection threshold.

5. Discussion and Future Research Directions

The experimental results of Section 4.1 suggest that the proposed scale-space methodology can be successful in detecting small changes in a hyperspectral image. To be useful in practice, such a method must have both high sensitivity and high specificity and our results for the artificial mole data clearly show promise in this respect. We are currently in the process of acquiring a large number of HSI data sets related to skin moles and lesions in collaboration with several hospitals in the Canary Islands, Spain. Our long term goal is to design a successful classifier for such data and the preliminary results obtained so far are promising [14]. However, we believe that a system capable of monitoring dynamical changes in a mole will be even more important as it is likely to be the best way to detect severe skin cancer at an early stage. In the future, we will therefore work on the development of such a system and our ultimate goal is to design a decision support tool based on just a few frequency bands so that an affordable version could be implemented on a smart phone and thereby be available for use on an individual basis.

One aspect of hyperspectral image data not utilized in the present study is its spatial structure. Taking spatial information into account is important because it can significantly improve the interpretation of the data when changes have been detected. Spatial information can be used both in the development of the change detection algorithm and in the interpretation of the results. Besides mole monitoring applications, a successful change detection method incorporating spatial information could perhaps also be used in the analysis of brain fMRI data for the detection of early signs of, for example, Alzheimer's disease [16].

Another area where the present methodology can be directly applied is in the design of robust controllers for Type 1 Diabetes patients. Successful results in this area are currently being obtained by using reinforcement learning (RL), see, for example, [17]. In the design of such machine learning algorithms, a good description of the patient's state space is needed for the algorithm to be able to learn better strategies. The state space contains information used to describe the patient's condition at a given time. Typically, the elements of state space in this context are time series of the most recent past blood glucose levels of the patient. At the beginning of the learning phase of an RL algorithm, the state space may be chosen reasonably coarse. During the learning process, the state space then may need to change because the algorithm encounters new states, that is, new glucose level time series, not included in the initial state space. The detection of such changes in the state space time series can be accomplished by the kind of methods discussed in this article. Research in this direction will therefore be pursued in the near future.

We also plan to further develop our approach to the analysis of fish freshness discussed in Section 4.2. For the design of a practical system that can be reliably used in fish industry one must first analyze data sets from several different fish at several time points after capturing. Then it is possible estimate both the within variance (of a day) and the between variances (between different time points) exhibited by the hyperspectral signatures. We will acquire such data in the future and the goal will be to perform an analysis that demonstrates how early changes in fish (or other types of food) quality can be detected in a reliable way.

Finally, we believe that the proposed methodology can be useful in combating problems in the so-called "$p > n$" problems now commonly found in statistical data analyses. Here p and n refer to the number of model parameters and the number of available observations, respectively, and such problems are very common in applications that involve high dimensional data, see e.g., [18,19]. The methodology developed in this article was partly motivated by the need to improve the technique of Hindberg et al. [7] which was originally designed exactly for the $p > n$ situation where common covariance matrix based multivariate methods such as PCA are useless. Being clearly an improvement of the technique of Hindberg et al., the method developed in this article is potentially useful in the analysis of such high dimensional data.

6. Concluding Remarks

We have developed a scale-space methodology that can successfully detect small changes in curve data. In addition, the developed methodology has the potential to produce few false alarms, an important feature for any detection method. We analyzed the performance of the proposed method on data with artificial and real changes. In addition, we compared the new method to some natural competitors and demonstrated that it at least in some cases outperforms them. Finally, we outlined several future research directions for the new methodology that can lead to important new findings.

Author Contributions: S.U.: Method and writing, T.H.J.: Method and writing, J.I.Z.: Method and review, S.O.: Method and review, L.H.: Method and review, G.M.C.: Method and review, H.F.: Method and review, F.G.: Conseptualization and review. All authors have read and agreed to the published version of the manuscript.

Funding: This project is supported by Tromsø Research Foundation through TFS project ID: 16_TF_FG.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Asokan, A.; Anitha, J. Change detection techniques for remote sensing applications: A survey. *Earth Sci. Inform.* **2019**, *12*, 143–160. [CrossRef]
2. Truong, C.; Oudre, L.; Vayatis, N. Selective review of offline change point detection methods. *Signal Process.* **2019**, *167*. [CrossRef]
3. Lu, G.; Fei, B. Medical hyperspectral imaging: A review. *J. Biomed. Opt.* **2014**, *19*. [CrossRef] [PubMed]
4. Alander, J.T.; Bochko, V.; Martinkauppi, B.; Saranwong, S.; Mantere, T. A Review of Optical Nondestructive Visual and Near-Infrared Methods for Food Quality and Safety. *Int. J. Spectrosc.* **2013**, *2013*. [CrossRef]
5. Olsen, L.R.; Sørbye, S.H.; Godtliebsen, F. A scale-space approach for detecting non-stationarities in time series. *Scand. J. Stat.* **2007**, *35*, 119–138. [CrossRef]
6. Holmström, L.; Pasanen, L. Bayesian scale space analysis of differences in images. *Technometrics* **2012**, *54*, 16–29. [CrossRef]
7. Hindberg, K.; Hannig, J.; Godtliebsen, F. A novel scale-space approach for multinormality testing and the k-sample problem in the high dimension low sample size scenario. *PLoS ONE* **2019**, *14*. [CrossRef] [PubMed]
8. Lindeberg, T. Scale-space theory: A basic tool for analyzing structures at different scales. *J. Appl. Stat.* **1994**, *21*, 225–270. [CrossRef]
9. Holmström, L.; Pasanen, L. Statistical Scale Space Methods. *Int. Stat. Rev.* **2017**, *85*, 1–30. [CrossRef]
10. Wand, M.P.; Jones, M.C. Kernel Smoothing. In *Monographs on Statistics & Applied Probability*; Chapman & Hall: London, UK, 1994; ISBN 9780412552700.
11. Hochberg, Y.; Tamhane, A.C. *Multiple Comparison Procedures*; Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics; Wiley: Hoboken, NJ, USA, 1987; ISBN 9780471822226.
12. Hannig, J.; Marron, J.S. Advanced distribution theory for SiZer. *J. Am. Stat. Assoc.* **2006**, *101*, 484–499. [CrossRef]
13. Lawrence, K.C.; Park, B.; Windham, W.R.; Mao, C. Calibration of a pushbroom hyperspectral imaging system for agricultural inspection. *Trans. Am. Soc. Agric.* **2003**, *46*, 513–521. [CrossRef]
14. Fabelo, H.; Melián, V.; Martínez, B.; Beltrán, P.; Ortega, S.; Marrero, M.; Callicó, G.M.; Sarmiento, R.; Castaño, I.; Carretero, G.; et al. Dermatologic Hyperspectral Imaging System for Skin Cancer Diagnosis Assistance. In Proceedings of the 2019 XXXIV Conference on Design of Circuits and Integrated Systems (DCIS), Bilbao, Spain, 20–22 November 2019.
15. Johnson, R.A.; Wichern, D.W. *Applied Multivariate Statistical Analysis (Classic Version)*; Pearson Modern Classics for Advanced Statistics Series; Pearson: London, UK, 2018; ISBN 9780134995397.
16. Malik, F.; Farhan, S.; Fahiem, M.A. An ensemble of classifiers based approach for prediction of Alzheimer's disease using fMRI images based on fusion of volumetric, textural and hemodynamic features. *Adv. Electr. Comput. Eng.* **2018**, *18*, 61–71. [CrossRef]
17. Tejedor, M.; Woldaregay, A.Z.; Godtliebsen, F. Reinforcement learning application in diabetes blood glucose control: A systematic review. *Artif. Intell. Med.* **2020**, *104*. [CrossRef]
18. Giraud, C. *Introduction to High-Dimensional Statistics*. In *Monographs on Statistics & Applied Probability*; Chapman & Hall: London, UK, 2014; ISBN 9781482237955.

19. Aoshima, M.; Shen, D.; Shen, H.; Yata, K.; Zhou, Y.H.; Marron, J.S. A survey of high dimension low sample size asymptotics. *Aust. New Zealand J. Stat.* **2018**, *60*, 4–19. [CrossRef] [PubMed]

Chapter 11.

Paper III

Towards detection and classification of microscopic foraminifera using transfer learning

Thomas Haugland Johansen, Steffen Aagaard Sørensen

Published

Towards detection and classification of microscopic foraminifera using transfer learning

Thomas Haugland Johansen*[1] and Steffen Aagaard Sørensen[2]

[1]Department of Mathematics and Statistics, UiT The Arctic University of Norway
[2]Department of Geology, UiT The Arctic University of Norway

Abstract

Foraminifera are single-celled marine organisms, which may have a planktic or benthic lifestyle. During their life cycle they construct shells consisting of one or more chambers, and these shells remain as fossils in marine sediments. Classifying and counting these fossils have become an important tool in e.g. oceanography and climatology. Currently the process of identifying and counting microfossils is performed manually using a microscope and is very time consuming. Developing methods to automate this process is therefore considered important across a range of research fields. The first steps towards developing a deep learning model that can detect and classify microscopic foraminifera are proposed. The proposed model is based on a VGG16 model that has been pretrained on the ImageNet dataset, and adapted to the foraminifera task using transfer learning. Additionally, a novel image dataset consisting of microscopic foraminifera and sediments from the Barents Sea region is introduced.

1 Introduction

Foraminifera are ubiquitous ocean dwelling single-celled microorganisms that may have a planktic (living in the water column) or benthic (living at or within the seabed) lifestyle. During their life cycle foraminifera construct shells with one or more chambers. The shells are commonly composed of calcium carbonate (calcareous foraminifera) or constructed from sediment particles cemented together (agglutinated foraminifera). They are recognizable due to their interspecies morphological differences. The shells remain in the marine sediments as fossils, and can be extracted from rock or marine sediment samples. Foraminifera are common in both modern and ancient environments and have become invaluable tools in oceanographic and geoscience research as well as in petroleum exploration. For example in paleo-research, fossilized foraminiferal fauna compositions and/or chemical composition of individual shells are frequently used to infer past changes in ocean temperature, salinity, ocean chemistry, and global ice volume [1,9,14]. In ecotoxicology and pollution monitoring studies, changes in foraminiferal abundance, morphology and faunal composition are used for detecting ecosystem contamination [6]. In the petroleum industry, foraminiferal analysis is an important tool to infer ages and paleoenvironments of sedimentary strata in oil wells during exploration, which aids the detection of potential hydrocarbon deposits [3,13].

Statistical counting of foraminifera species, their number and distribution, represents important data for marine geological climate and environmental research and in petroleum exploration. Counting, identification and picking of foraminifera in prepared sediment samples using a microscope is a very time and resource demanding process, which has practically been conducted the same way since the use of microscope foraminiferal studies started in the early 1800's. Progress in deep learning makes it possible to automate this work, which will contribute to better quality, higher quantity, reduced resource usage, and more cost effective data collection. Existing research groups have already started with image recognition of foraminifera [4,8,11,17], but the training data currently needs to be "tailor made" with the most abundant foraminiferal species for a specific geographical region.

*Corresponding Author: thomas.h.johansen@uit.no

https://doi.org/10.7557/18.5144

2 Transfer learning

There are a number of transfer learning methods used in deep learning, and in the presented experiments two such methods are implemented, namely feature extraction and fine tuning.

The strengths of a deep convolutional neural network (CNN) model is its many layers of filters, learned by training on millions of images [12]. Learning the weights of these layers can require an enormous amount of images, depending on e.g. the depth and complexity the model, the input domain, etc. However, the learned filters represent somewhat abstract feature detectors that can be transferred to new domains [2, 16]. In other words, it is possible to re-use the weights of a pretrained CNN model for new classification tasks. In its simplest form this is achieved by using the convolutional blocks of the model as a feature extractor, and the extracted features can then be passed to any classifier. The weights of the classifier need to be learned, but the weights of the pretrained filter layers are preserved or "frozen". Typically the classifier is chosen such that it performs well at the task of predicting output labels using the extracted features, while also being tractable to train.

It is also possible to re-train some layers of the CNN to optimize the extracted features to the new domain, which is referred to as fine tuning. This will then be a trade-off between adapting the pretrained model to the new image modalities, but with the risk of overfitting given the typically small size of the training dataset. Which layers to re-train typically depend on several factors, such as similarity between the new and the original image modalities.

3 Monte Carlo dropout

The complexity of a CNN classifier makes the output inconceivable in terms of the usual image feature interpretation, and there is a need for a measure of uncertainty. A step in that direction is to allow for stochastic prediction through Monte Carlo dropout.

Dropout is a regularization technique frequently used when training deep neural network models to reduce the chance of overfitting [15]. The basic idea is that a specified percentage of weights for some layers in the model are set to zero, effectively turning off the corresponding units in that layer. This percentage is referred to as the dropout rate and is considered a model hyperparameter. Which units to drop during training are chosen at random, typically sampling from a uniform distribution. One intuition behind dropout is that it helps prevent units from co-adapting, which might otherwise lead to "memorization" of training data. See Figure 1 for an illustrative toy example of how dropout behaves with a rate of 50%.

Once the model has been trained, the dropout rate is normally set to zero to ensure predictions are deterministic. Since units are dropped at random, predictions are stochastic, and this is the underlying idea of Monte Carlo dropout [7]. By considering dropout to be a Bayesian approximator in some sense, it becomes possible to analyze e.g. model uncertainty.

Assume a neural network f with model parameters $\mathbf{W}$ has been trained such that

$$\tilde{\mathbf{Y}} = f(\mathbf{X}; \mathbf{W}), \tag{1}$$

where $\tilde{\mathbf{Y}}$ is the predicted output for some dataset $\mathbf{X}$ with true output $\mathbf{Y}$. Monte Carlo dropout can then be implemented by iterating over the dataset N times collecting the output predictions,

$$\tilde{\mathbf{Y}}_i = f(\mathbf{X}; \mathbf{W}_i), \quad i = 1, \ldots, N \tag{2}$$

where $\mathbf{W}_i$ represents the model parameters for the i-th iteration after applying dropout. Using the collected predictions, Monte Carlo estimates of the predictive mean and variance can be computed,

$$\tilde{\boldsymbol{\mu}} = \frac{1}{N} \sum_{i=1}^{N} \tilde{\mathbf{Y}}_i, \tag{3}$$

$$\tilde{\boldsymbol{\sigma}} = \frac{1}{N} \sum_{i=1}^{N} \left(\tilde{\mathbf{Y}}_i - \tilde{\boldsymbol{\mu}} \right)^2. \tag{4}$$

The predictive mean $\tilde{\boldsymbol{\mu}}$ can be interpreted as the ensemble prediction for N different models. Similarly, the uncertainty of the ensemble predictions can be expressed using the predictive variance.

4 Preparing the datasets

The materials (foraminifera and sediment) used for the present study were collected from sediment

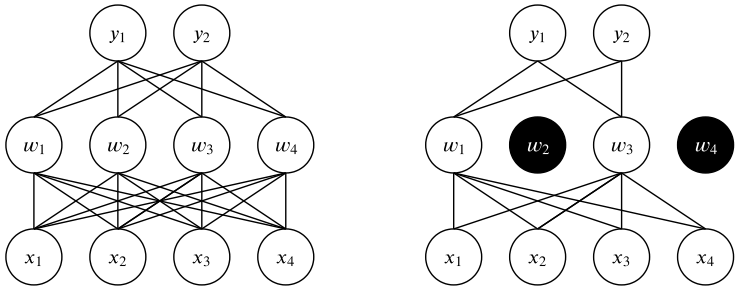


Figure 1: Toy example illustrating a neural network with and without dropout applied.

cores retrieved in the Arctic Barents Sea region. In order to achieve a good representation of the planktic and benthic foraminiferal fauna of the area, the specimens were picked from sediments influenced by Atlantic, Arctic, polar, and coastal waters representing different ecological environments. Foraminiferal specimens (planktics, benthics, agglutinated benthics) were picked from the 100 µm to 1000 µm size fraction of freeze dried and subsequently wet sieved sediments. Sediment grains representing a common sediment matrix were also sampled from the 100 µm to 1000 µm size range. The basis for the datasets were collected by photographing either pure benthic (calcareous or agglutinated), planktic assemblages, or sediments containing no foraminiferal specimens. In other words, each image contained only specimens belonging to one of four high-level classes; planktic, calcareous benthic, agglutinated benthic, sediment. This approach simplified the task of labeling each individual specimen with the correct class. All images were captured with a 5 megapixel Leica DFC450 digital camera mounted on a Leica microscope.

From each of the images collected from the microscope, smaller images of each individual specimen were extracted using a very simple, yet effective, object detection scheme based on Gaussian filtering, grayscale thresholding, binary masking and connected components. The first pass of Gaussian filtering, grayscale thresholding and binary masking was tuned to remove the metallic border present in each image, which can be seen in Figure 2. The next pass of filtering, thresholding and masking was tuned to detect the foraminifera and sediment candidates. Very small objects, which included remnant particulates (considered noise) from e.g. damaged specimens, were discarded based on the number of connected components; all candidates with less than 1024 pixels were discarded. After selecting candidates from the original microscope images, all of the individual specimen images were extracted by placing a 224×224 pixel crop region at the "center of mass" of each candidate. An example from this process can be seen in Figure 2.

Upon completing the object detection and image extraction procedure, the result was a dataset containing a total of 2673 images. These images were then stratified into training, validation and test sets using a 80/10/10 split. Examples of extracted images can be seen in Figure 3.

5 Experiments

All experiments presented are based on a VGG16 [12] model that had been pretrained on the ImageNet [5] dataset. The choice of model was made primarily due to prior experience and familiarity with the architecture.

5.1 Model design and training

Using a pretrained VGG16 model, feature vectors were extracted from each of the foraminifera and sediment images in the dataset. See Figure 4 for a simplified illustration of the VGG16 model architecture. The feature extraction procedure was done by removing the fully-connected dense layers, the so called "classification head", at the end of the VGG16 model. Feature vectors were then extracted from the last convolutional block, and used as input features to a new deep neural network model designed to classify foraminifera and sediment. This new classification model went through several designs during initial prototyping, varying in number

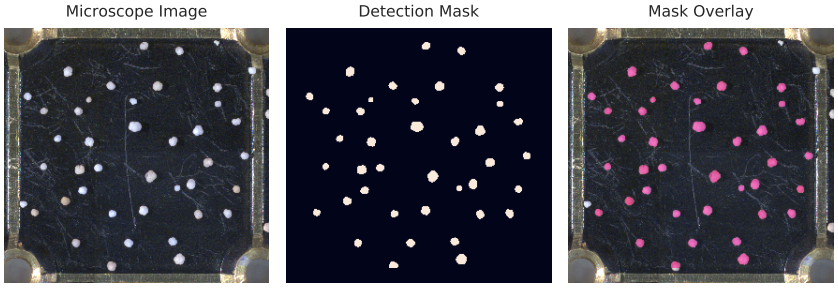


Figure 2: Examples from the detection and extraction procedure used to create the foraminifera dataset.

Layer Type	Input Dim.	Output Dim.
VGG16	224×224×3	7×7×512
Dense (ReLU)	25088	512
Dense (ReLU)	512	64
Dense (Softmax)	64	4

Table 1: High-level summary of the deep learning model used to classify foraminifera and sediments.

of layers and units per layer. Ultimately, hyperparameter tuning was performed to finalize the design of the classifier. This was done using a grid search approach, which tested 72 different permutations of units per layer, dropout rate, and optimization algorithm. The final end-to-end model architecture can be seen summarized in Table 1.

The model was first trained with all weights for the VGG16 model being fixed, and thus only the weights of the new classification head were optimized. All training was done using a batch size of 32, cross entropy loss, and an *Adam* [10] optimizer with an initial learning rate of 10^{-4}. To reduce the chance of overfitting, early stopping was implemented based on the validation accuracy computed at the end of each training epoch. On average, due to early stopping, each training session stopped after 7 epochs, with each epoch consisting of 260 training steps. After initial training of the classification model on feature vectors extracted from the VGG16 model, fine-tuning was implemented to improve classification accuracy. This was achieved by "unfreezing" the last two convolutional blocks of the VGG16 model, thus allowing the model to specialize those parameters to the new classification task. The initial learning rate during fine-tuning was reduced to 10^{-7} to ensure smaller, incremental gradient updates.

Given the relatively small dataset, image augmentation was implemented to synthetically boost the number of training images. The augmentations consisted of flipping, rotating, as well as changing brightness, contrast, hue, and saturation. Flipping was done horizontally, and rotations in increments of 90 degrees. Brightness, contrast and saturation values were randomly augmented by ±10%, whereas hue was augmented by ±5%. These augmentations were chosen based on qualitative analysis of the dataset to ensure they were both representative and valid. Each augmentation was applied in a randomized fashion to every image in a batch, each time a training batch was sampled.

The training procedure was repeated multiple times to reduce the effects of random initialization of model weights. After only training the classification head, the mean accuracy on the test data was $97.0 \pm 0.6\%$. Fine-tuning improved the results to a mean accuracy of $98.8 \pm 0.2\%$.

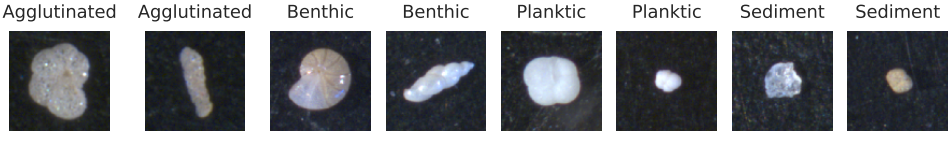


Figure 3: Examples of typical specimens from each of the four categories found in the image dataset.

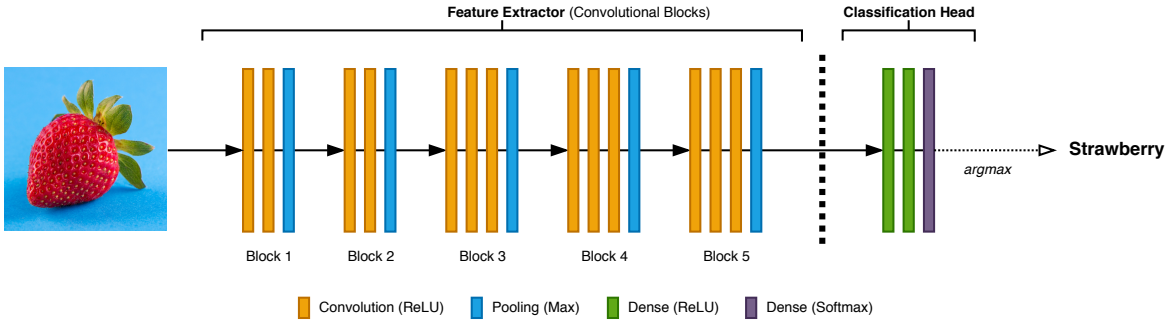


Figure 4: Simplified architecture diagram of the VGG16 model. Input images are passed through the convolutional blocks, and feature vectors are then transformed by dense layers into softmax predictions.

5.2 Model analysis

After training, Monte Carlo dropout was implemented in order to investigate and analyze the trained models. Model predictions were collected as expressed in (2) for $N = 100$, with all dropout layers turned on and using the entire test set. Predictive mean and variance were calculated using (3) and (4), respectively.

Using these results made it possible to uncover difficult cases in the dataset where the model was having problems with the classification. There were two scenarios; the model was uncertain about the prediction, or it was certain, but the prediction was incorrect. When studied qualitatively, some of the challenging images contained overexposed specimens that were missing details such as texture. In other cases, specimens were oriented in such a way that the morphological characteristics of the foraminifera were not visible. An example of an overexposed specimen can be seen in Figure 5. Some of the challenging cases were shown to a trained expert, which was able to correctly classify all specimens.

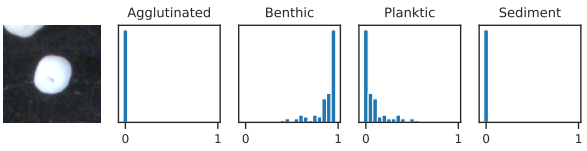


Figure 5: Overexposed planktic foraminifera, misclassified as benthic. Histograms represent distributions of softmax predictions from MC dropout.

The mean accuracy for all Monte Carlo simulations was $97.9 \pm 0.5\%$. Furthermore, by considering each simulation to be part of an ensemble of models with a majority voting scheme, the accuracy of the ensemble predictions was 98.5%. These results are comparable to the model without Monte Carlo dropout.

6 Concluding remarks

Based on the presented experiments it is clear that training deep learning models to accurately classify microscopic foraminifera is possible. Using VGG16 pretrained on ImageNet to extract features from foraminifera produces very promising results, which can then be further improved by fine-tuning the pretrained model. The results are comparable to equivalent efforts by other research group using different datasets of foraminifera and sediments.

To uncover images in the dataset that the model is uncertain about techniques such as Monte Carlo dropout can used. These results can then be used to identify classes that need more training data, or perhaps alludes to further image augmentation, etc.

Future work should involve investigations using model architectures other than VGG16 should be conducted, comparing differences in prediction accuracy, computational efficiency during training and inference, and so forth. Once bigger datasets become available, efforts should also invested towards training novel models from scratch, and comparing to pretrained models.

References

[1] S. Aagaard-Sørensen, K. Husum, K. Werner, R. F. Spielhagen, M. Hald, and T. M. Marchitto. A late glacial–early holocene multiproxy record from the eastern fram strait, polar north atlantic. *Marine Geology*, 355:15–26, 2014.

[2] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[3] R. S. Boardman, A. H. Cheetham, and A. J. Rowell. *Fossil invertebrates*. Blackwell Scientific Publications, 1987.

[4] T. de Garidel-Thoron, R. Marchant, E. Soto, Y. Gally, L. Beaufort, C. T. Bolton, M. Bouslama, L. Licari, J.-C. Mazur, J.-M. Brutti, et al. Automatic picking of foraminifera: Design of the foraminifera image recognition and sorting tool (first) prototype and results of the image classification scheme. In *AGU Fall Meeting Abstracts*, 2017.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009.

[6] F. Frontalini and R. Coccioni. Benthic foraminifera as bioindicators of pollution: a review of italian research over the last three decades. *Revue de micropaléontologie*, 54(2):115–127, 2011.

[7] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059, 2016.

[8] Q. Ge, B. Zhong, B. Kanakiya, R. Mitra, T. Marchitto, and E. Lobaton. Coarse-to-fine foraminifera image segmentation through 3d and deep features. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–8. IEEE, 2017.

[9] M. Hald, C. Andersson, H. Ebbesen, E. Jansen, D. Klitgaard-Kristensen, B. Risebrobakken, G. R. Salomonsen, M. Sarnthein, H. P. Sejrup, and R. J. Telford. Variations in temperature and extent of atlantic water in the northern north atlantic during the holocene. *Quaternary Science Reviews*, 26(25-28):3423–3440, 2007.

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

[11] R. Mitra, T. Marchitto, Q. Ge, B. Zhong, B. Kanakiya, M. Cook, J. Fehrenbacher, J. Ortiz, A. Tripati, and E. Lobaton. Automated species-level identification of planktic foraminifera using convolutional neural networks, with comparison to human performance. *Marine Micropaleontology*, 147:16–24, 2019.

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

[13] A. Singh. Micropaleontology in petroleum exploration. In *7th International Conference and Exposition of Petroleum Geophysics*, pages 14–16, 2008.

[14] R. F. Spielhagen, K. Werner, S. A. Sørensen, K. Zamelczyk, E. Kandiano, G. Budeus, K. Husum, T. M. Marchitto, and M. Hald. Enhanced modern heat transfer to the arctic by warm atlantic water. *Science*, 331(6016):450–453, 2011.

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958, 2014.

[16] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *NeurIPS*, pages 3320–3328, 2014.

[17] B. Zhong, Q. Ge, B. Kanakiya, R. M. T. Marchitto, and E. Lobaton. A comparative study of image classification algorithms for foraminifera identification. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–8. IEEE, 2017.

Chapter 12.

Paper IV

Instance Segmentation of Microscopic Foraminifera

Thomas Haugland Johansen, Steffen Aagaard Sørensen, Kajsa Møllersen, Fred Godtliebsen

In submission

Article

Instance Segmentation of Microscopic Foraminifera

Thomas Haugland Johansen [1,*], Steffen Aagaard Sørensen [2], Kajsa Møllersen [3], Fred Godtliebsen [1]

[1] Department of Mathematics and Statistics, UiT The Arctic University of Norway
[2] Department of Geosciences, UiT The Arctic University of Norway
[3] Department of Community Medicine, UiT The Arctic University of Norway
* Correspondence: thomas.h.johansen@uit.no

Abstract: Foraminifera are single-celled marine organisms that construct shells that remain as fossils in the marine sediments. Classifying and counting these fossils are important in e.g. paleo-oceanographic and -climatological research. However, the identification and counting process has been performed manually since the 1800s and is laborious and time-consuming. In this work, we present a deep learning-based instance segmentation model for classifying, detecting, and segmenting microscopic foraminifera. Our model is based on the Mask R-CNN architecture, using model weight parameters that have learned on the COCO detection dataset. We use a fine-tuning approach to adapt the parameters on a novel object detection dataset of more than 7000 microscopic foraminifera and sediment grains. The model achieves a (COCO-style) average precision of 0.78 ± 0.00 on the classification and detection task, and 0.80 ± 0.00 on the segmentation task. When the model is evaluated without challenging sediment grain images, the average precision for both tasks increases to 0.84 ± 0.00 and 0.86 ± 0.00,, respectively. Predictions results are analyzed both quantitatively and qualitatively and discussed. Based on our findings we propose several directions for future work, and conclude that our proposed model is an import step towards automating the identification and counting of microscopic foraminifera.

Keywords: foraminifera; instance segmentation; object detection; deep learning

1. Introduction

Foraminifera are microscopic (typically smaller than 1 mm) single-celled marine organisms (protists) that during their life cycle construct shells from various materials that readily fossilize in sediments and can be extracted and examined. Roughly 50 000 species have been recorded of which approximately 9 000 are living today [1]. Foraminiferal shells are abundant in both modern and ancient sediments. Establishing the foraminifera faunal composition and distribution in sediments, and measuring the stable isotopic and trace element composition of shell material have been effective techniques for reconstructing past ocean and climate conditions [2–4]. Foraminifera have also proven valuable as bio-indicators for anthropogenically introduced stress to the marine environment [5]. After a sediment core has been retrieved from the seabed, a range of procedures are performed in the laboratory before the foraminiferal specimens can be identified and extracted under the microscope by a geoscientist using a brush or needle. From each core, several layers are extracted, and each layer is regarded as a sample. To establish a statistically robust representation of the fauna, 300-500 specimens are identified and extracted per sample. The time-consumption of this task is 2–8 hours/sample, depending on the complexity of the sample and the experience level of the geoscientist. A typical study consists of 100-200 samples from one or several cores, and the overall time-consumption in just identifying the specimens is vast. Recently developed deep learning models show promising results towards automating parts of the identification and extraction process [6–10].

Figure 1 shows an example of a prepared foraminifera sample — microscopic objects spread out on a plate and photographed through a microscope. Of particular interest is the classification of each object into high-level foraminifera classes, which then serves as input for estimation of the environment in which sediment was produced. This task consists of identifying relevant objects, particularly separate sediment from foraminifera, and recognize foraminifera classes based on shapes and structures of each object.

Citation: Johansen, T. H.; Sørensen, S. A.; Møllersen, K.; Godtliebsen, F. Instance Segmentation of Microscopic Foraminifera. *Preprints* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

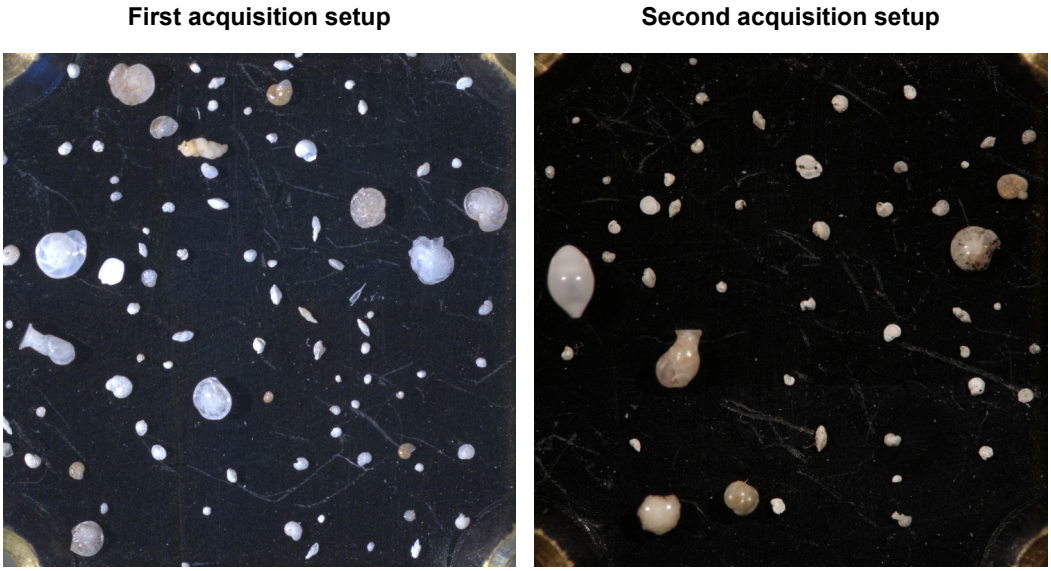


Figure 1. Examples of images from the two different image acquisition setups used during three data collection phases. **Left**: Calcareous benthics photographed with the first acquisition setup used during the first data collection phase. **Right**: Calcareous benthics photographed with the second acquisition setup used during the second and third data collection phases.

Object classification in images is one of the great successes of deep learning, and conquer new applications as new methods are developed and high-quality data are made available for training and testing. In a deep learning context, a core task of object classification is instance segmentation. Not only must the objects be separated from the background, but the objects themselves must be separated from each other, so that adjacent objects are identified, and not treated as one single object.

Automatic foraminifera identification has great practical potential: the time saved for highly qualified personnel is substantial; it is the overall proportion of foraminifera classes that is the primary interest, and it is therefore robust to the occasional misclassification; and availability of deep learning algorithms that integrates object detection, instance segmentation and object classification. The lack of publicly available data sets for this particular deep learning application has been an obstacle, but with a curated private data set, soon-to-be published, the stars have finally aligned for an investigation into the potential of applying deep learning to foraminifera classification.

The manuscript is organized as follows:
In Section 2.1, we describe the acquisition and preparation of the dataset, and its final attributes. In Section 2.2, we give an overview of the Mask R-CNN model applied to foraminifera images. In Section 2.3, we give a detailed description of the experimental setup. To present the results, we have chosen to include training behavior (Section 3.1), since this is a first attempt for foraminifera application. Further, we give a detailed presentation of the performance from various aspects and different thresholds (Section 3.2) for a comprehensive understanding of strengths and weaknesses. Section 4 then emphasizes and discusses the most interesting findings, both in terms of promising performance and in terms of future work (Section 4.1). We round off with a conclusion (Section 5) to condense the discussion into three short statements.

2. Materials and Methods

The work presented in this article was performed in two distinct phases; first a novel object detection dataset of microscopic foraminifera was created, and then a pretrained Mask R-CNN model [11] was adapted and fine-tuned on the dataset.

2.1. Dataset Curation

All presented materials (foraminifera and sediment) were collected from sediment cores retrieved in the Arctic Barents Sea region. The specimens were picked from sediments influenced by Atlantic, Arctic, polar, and coastal waters representing different ecological environments. This was done to ensure good representation of the planktic and benthic foraminiferal fauna in the region. Foraminiferal specimens (planktics, benthics, agglutinated benthics) were picked from the 100 µm to 1000 µm size fraction of freeze dried and subsequently wet sieved sediments. Sediment grains representing a common sediment matrix were also sampled from the 100 µm to 1000 µm size range.

The materials were prepared and photographed at three different points in time, with two slightly different image acquisition systems. During the first two rounds of acquisition, every image contains either pure benthic (agglutinated or calcareous), planktic assemblages, or sediment grains containing no foraminiferal specimens. In other words, each image contained only specimens belonging to one of four high-level classes; agglutinated benthic, calcareous benthic, planktic, sediment grain. This approach greatly simplified the task of labeling each individual specimen with the correct class. In order to better mimic a real-world setting with mixed objects, the third acquisition only contained images where there was a realistic mixture of the four object classes. To get the necessary level of magnification and detail, four overlapping images were captured from the plates on which the specimens were placed, where each image corresponded to a distinct quadrant of the plate. The final images were produced by stitching together the mosaic of the four partially overlapping images.

All images from the first acquisition were captured with a 5 megapixel Leica DFC450 digital camera mounted on a Leica Z16 APO fully apochromatic zoom system. The remaining two acquisitions were captured using a 51 megapixel Canon EOS 5DS R camera mounted on a Leica M420 microscope. The same Leica CLS 150x (twin goose-neck combination light guide) was used for all acquisitions, but with slightly different settings. The light power was set to 4 and 4 for the first acquisition, and to 3 and 3 for the remaining two. No illumination or color correction was performed, in an attempt to mimic a real-world scenario of directly detecting, classifying and segmenting foraminifera placed under a microscope. Examples of the differences in illumination settings can be seen in Figure 1.

To create the ground truth, a simple, yet effective, hand-crafted object detection pipeline [10] was ran on each image, which produced initial segmentation mask candidates. The pipeline consisted of two steps of Gaussian smoothing, then grayscale thresholding followed by a connected components approach to detect individual specimens. Some parameters such as the width of Gaussian filter kernel, and threshold levels, were hand-tuned to produce good results for each image in the dataset. A simple illustration of the preprocessing pipeline can be seen in Figure 2. For full details, see Johansen and Sørensen [10].

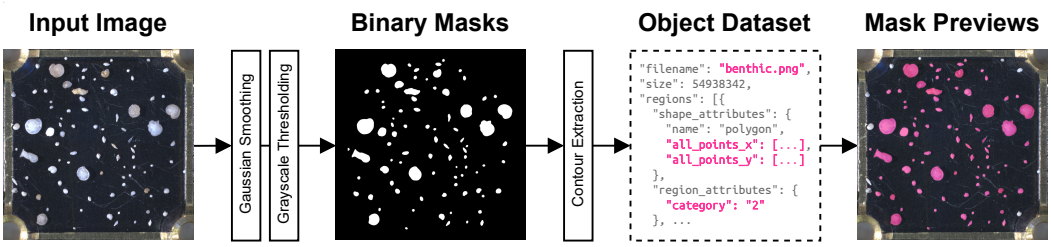


Figure 2. High-level summary of the dataset creation pipeline.

After obtaining the initial segmentation mask dataset, all masks were manually verified and adjusted using the VGG Image Annotator [12,13] software. Additionally, approximately 2000 segmentation masks were manually created (using the same software) for objects not detected by the detection pipeline. The end result is a novel object detection

dataset consisting of 104 images containing over 7000 segmented objects. Full details on the final dataset can be found in Table 1.

Table 1. Detailed breakdown of the object dataset, where row holds information for a specific microscope image acquisition phase. The first and second phases contain only "pure" images where every object is of a single class, whereas the third phase images contain only mixtures of several classes.

			Objects per class			
Phase	**Images**	**Objects**	**Agglutinated**	**Benthic**	**Planktic**	**Sediment**
First	48	3775	172	897	726	1980
Second	41	2604	583	695	657	669
Third	15	633	154	156	155	168
Combined	104	7012	909	1748	1538	2817

2.2. Instance Segmentation using Deep Learning

Mask R-CNN [11] is a proposal-based deep learning framework for instance segmentation, and is an extension to Fast/Faster R-CNN [14,15]. In the Fast/Faster R-CNN framework the model has two output branches, one that performs bounding box regression and another that performs classification. The input to these two branches are pooled regions of interest (RoIs) produced from features extracted by a convolutional neural network (CNN) backbone. This is extended in Mask R-CNN by adding an extra (decoupled) output branch, which predicts segmentation masks on each RoI. Figure 3 shows a simple, high-level representation of the Mask R-CNN model architecture. Several alternatives to Mask R-CNN exist, such as PANet [16], TensorMask [17], CenterMask [18], and SOLOv2 [19]. We chose to use the Mask R-CNN framework for two key reasons: (1) it predicts bounding boxes, class labels, and segmentation masks at the same time in a single forward-pass, and (2) pretrained model parameters are readily available, removing the need to train the model from scratch.

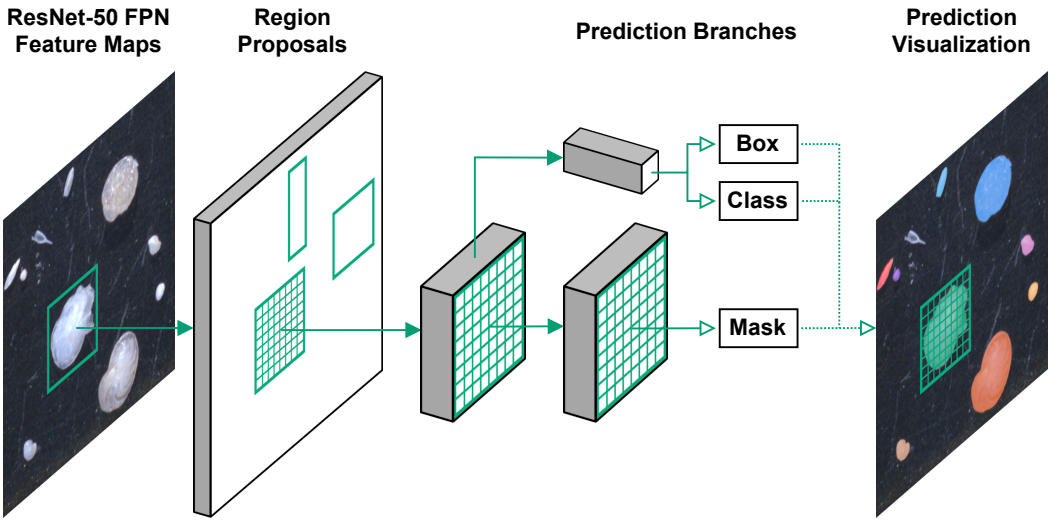


Figure 3. Simple, sketch-like depiction of the Mask R-CNN model architecture.

Due to its flexible architecture, there are numerous ways to design the feature extraction backbone of a Mask R-CNN model. We chose a model design based on a ResNet-50 [20] Feature Pyramid Network (FPN) [21] backbone for feature extraction and RoI proposals. To avoid having to train the model from scratch, we applied model parameters pretrained on the COCO dataset [22]. The object detection model and all experiment code was im-

plemented using Python 3.8, PyTorch 1.7.1, and torchvision 0.8.2. The pretrained model weights were downloaded via the torchvision library.

2.3. Experiment Setup and Training Details

The original Mask R-CNN model was trained using 8 GPUs and a batch size of 16 images, with 2 images per GPU. We did not have access to that kind of compute resources, and were instead limited to a single NVIDIA TITAN Xp GPU, which also meant our training batches only consisted of a single image. The end result of this was slightly more unstable loss terms and gradients, so we carefully tested many different optimization methods, learning rates, learning rate scheduling, and so on.

The dataset was split (with class-level stratification) into separate training and test sets, using a 2.47 : 1 ratio, which produced 74 training images and 30 test images. The training and test sets remained the same for all experiments. During training images were randomly augmented, which included horizontal and vertical flipping, brightness, contrast, saturation, hue, and gamma adjustments. Both the horizontal and vertical flips were applied independently, with a flip probability of 50% for both cases. Brightness and contrast factors were randomly sampled from $[0.9, 1.1]$, the saturation factor from $[0.99, 1.01]$, and hue from $[-0.01, 0.01]$. For the random gamma augmentation, the gamma exponent was randomly sampled from $[0.8, 1.2]$.

We ran the initial experiments using the Stochastic Gradient Descent (SGD) optimization method with Nesterov momentum [23] and weight decay. The learning rates tested were $\{10^{-3}, 5\times10^{-3}, 10^{-5}\}$, and the momentum parameter was set 0.9. For weight decay, we tested the values $\{0, 10^{-4}, 10^{-5}, 5\times10^{-5}\}$. In some experiments the learning rate was reduced by a factor of 10 after either 15 or 25 epochs. Training was stopped after 50 epochs. After the initial experiments with SGD we tested the Adam [24] optimization method. We tested the learning rates $\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. The weight decay parameter values were $\{0, 10^{-4}, 10^{-6}\}$. We used the same scheduled learning rate decay as with SGD for the initial experiment, and training was stopped after 50 epochs.

From on our initial experiments with SGD and Adam, we saw that the latter gave more stable loss terms during training and decided to use this method for the subsequent experiments. However, we experimented with a recent variant of the Adam optimizer with decoupled weight decay, referred to as AdamW [25]. We implemented a slightly adjusted scheduled learning rate decay, with a factor of 10 reduction after both 25 and 45 epochs of training. Because we used model parameters pretrained on the COCO dataset, we also ran experiments with fine-tuning the backbone model to adapt it to our target domain. For the fine-tuning experiments we tested when to "freeze"and "unfreeze" the backbone model parameters, i.e. when to fine-tune the backbone, as well as which layers of the backbone to fine-tune.

Based on all probing experiments and the hyperparameter tuning, our final model was trained using AdamW for 50 epochs. During the first 25 epochs of training the last three ResNet-50 backbone layers were fine-tuned, and then subsequently frozen. The initial learning rate was set to 10^{-5} and was reduced to 10^{-6} after 25 epochs, and further reduced to 10^{-7} after 45 epochs. We set the weight decay parameter to 10^{-4}. Using this configuration, we trained the model 10 times using different random number generator states to ensure valid results and to measure the robustness of the model.

3. Results

Model performance is evaluated using the standard COCO metrics for object detection and instance segmentation [26]. Specifically, we are using the average precision (AP) and average recall (AR) metrics averaged over 10 intersection-over-union (IoU) thresholds[1]and all classes. We also use the more traditional definition of AP, which is evaluated at a specific IoU, e.g. AP_{50} denotes the AP evaluated with an IoU of 0.5. Additionally, we

[1] The IoU thresholds range from 0.5 to 0.95 in increments of 0.05.

present conventional precision-recall curves with different evaluation configurations, e.g. per-class, per-IoU, and so forth. All presented precision and recall results were produced by evaluating models on the test split of the dataset.

3.1. Model Training

During training, all training losses were carefully monitored and reported both per-batch and per-epoch. Four of the key loss terms for the Mask R-CNN model can be seen in Figure 4, where each curve represents one of the 10 repeated training runs, with different initial random state. At the end of every training epoch, we evaluated the model performance in terms of the AP metric for both the detection and segmentation task on the test images. The per-epoch results for all 10 runs can be seen in Figure 5.

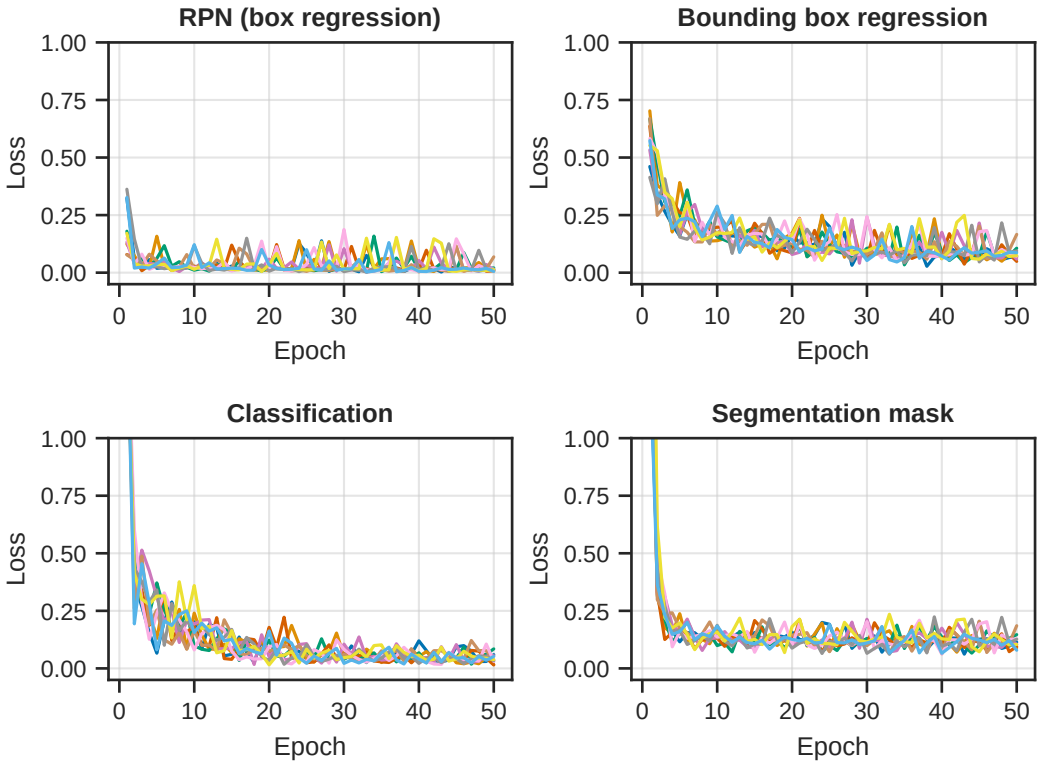


Figure 4. Evolution of the individual loss terms for each of the 10 training runs. **Top left**: The loss term for the RPN box regression sub-task. Fairly rapid convergence, but we can see the effect of the single-image batches in the curves. **Top right**: Bounding box regression loss for the detection branch. The convergence is slower when compared to the RPN loss, and perhaps slightly less stable. **Bottom left**: Object classification loss for the detection branch. Similar story can be seen here as with the bounding box regression, which suggests a possible challenge with the detection branch. **Bottom right**: The segmentation mask loss for the segmentation branch. Fast convergence, but again we see the effect of the single-image training batches.

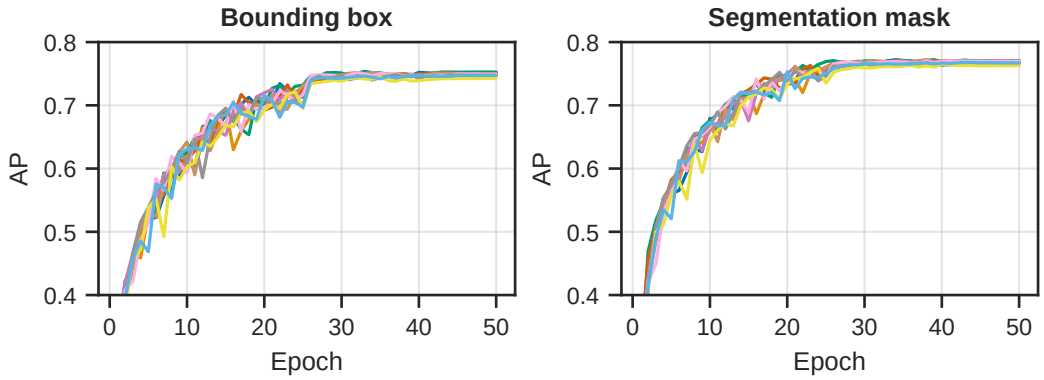


Figure 5. Evolution of the AP for both the detection and segmentation task, for each of the 10 runs, per epoch of training. Note that these results are based on evaluations using a maximum of 100 detections per image. **Left**: The AP for the detection (bounding box) task. A plateau is reached after about 30 epochs. **Right**: The AP for the segmentation mask task. We observe that the same type of plateau is reached here as with the detection task.

These results indicate that even though we reached some kind of plateau during training, we did not end up overfitting or otherwise hurt the performance on the test dataset. The AP for both tasks also reach a plateau, which is almost identical for all of the learned model parameters. This suggests that the training runs reached an upper limit on performance given the dataset, model design, and hyperparameters.

3.2. Evaluating the Model Performance

After the 10 training runs had concluded, we evaluated each model on the test data using their respective parameters from the final training epoch. Note that all precision and recall evaluations presented from this point onward are based on a maximum of 256 detections per image[2]. The mean AP across all of the 10 run was evaluated as 0.78 ± 0.00 for the detection task, and 0.80 ± 0.00 for the segmentation task. Table 2 shows a summary of the AP and AR metrics for both tasks, where each result is the mean and standard deviation of all training runs. Note that this table shows results averaged over all four classes, and also with the "sediment" class omitted from each respective evaluation, which will be discussed later.

Table 2. AP and AR scores for different IoU thresholds, evaluated with all object classes being considered and with the "sediment" class excluded.

	All classes		**Sans "sediment" class**	
	Bound. box	**Segm. mask**	**Bound. box**	**Segm. mask**
AP_{50}	0.90 ± 0.00	0.90 ± 0.00	0.94 ± 0.00	0.95 ± 0.00
AP_{75}	0.88 ± 0.00	0.90 ± 0.00	0.94 ± 0.00	0.94 ± 0.00
AP	0.78 ± 0.00	0.80 ± 0.00	0.84 ± 0.00	0.86 ± 0.00
AR	0.83 ± 0.00	0.84 ± 0.00	0.89 ± 0.00	0.90 ± 0.00

The precision-recall curves computed by averaging over all 10 training runs can be seen in Figure 6. Note the sharp and sudden drop in the curve around the recall threshold of 0.75, for both tasks.

[2] During training we ran model evaluations with a maximum of 100 detections per image due to lower computational costs and faster evaluation.

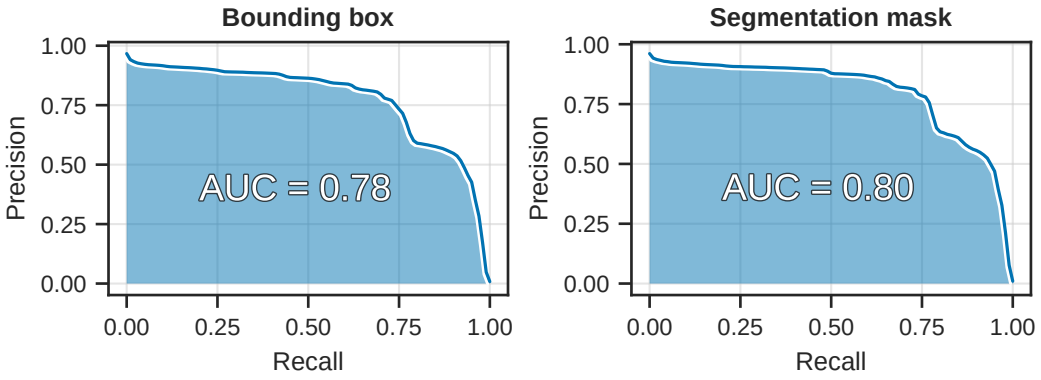


Figure 6. The mean average precision-recall curves for the 10 training runs, for both the detection and segmentation tasks. The area-under-the-curve (AUC) shown here is the same as our definition of AP, which can be seen in Table 2.

In order to investigate the sharp drop in precision and recall, we computed per-class precision and recall; the results can be seen in Figure 7. From the curves in the figure it is clear that the model is finding the "sediment" class particularly challenging. Notice how the precision rapidly goes towards zero slightly after the recall threshold of 0.75.

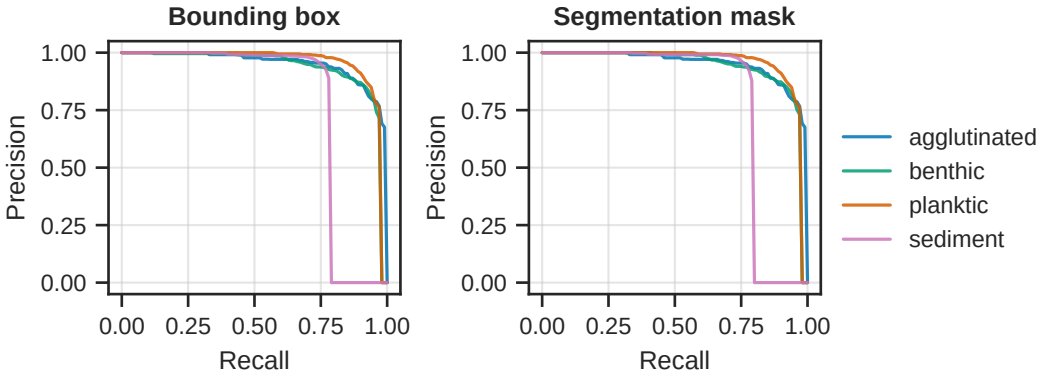


Figure 7. Precision-recall curves for each of the four object classes. **Left**: Per-class curves for the detection (bounding box) task. Performance is approximately the same for the "agglutinated", "benthic", and "planktic" classes, but is significantly worse for the "sediment" class. **Right**: The per-class curves for the segmentation task, which tells the same story as for the detection task.

We also wanted to determine how well the model performed at different IoU thresholds, so precision and recall were evaluated for the IoU thresholds $\{0.5, 0.75, 0.85, 0.95\}$. Figure 8 shows the precision-recall curves for all object classes, and Figure 9 shows the sans "sediment" class curves. From these results, it is clear that the model performs quite well at IoU thresholds up to and including 0.85, but at 0.95 the model does not perform well.

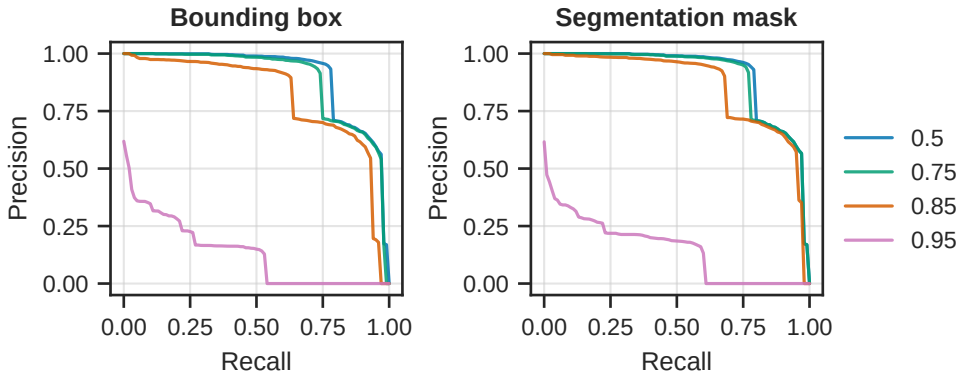


Figure 8. Precision-recall curves at different IoU thresholds, where each curve is based on the average for all four object classes. **Left**: PR curves for the detection (bounding box) task. There is a sharp drop in precision at the approximate recall thresholds $\{0.65, 0.74, 0.8\}$, which corresponds to the lower precision of the "sediment" class. **Right:** The same drop in precision is observer for the predicted masks, which can again be explained by the performance on the "sediment" class.

Based on the per-class and per-IoU results, it became evident that some test images containing only "sediment" class objects, were particularly challenging. This can in part be explained by the object density in these images, with multiple objects sometimes overlapping or casting shadows on each other. In the COCO context, these types of object clusters are referred to as a "crowd", and receive special treatment during evaluation. Importantly, none of the objects in our dataset have been annotated as being part of a "crowd".[3] Some examples of these dense object clusters can be seen in Figure A2 and Figure A3. By removing the "sediment" class from the evaluation, the AP score for the bounding box increased to 0.84, and for instance segmentation it increased to 0.86. Recall also increased significantly, which means that more target objects were correctly detected and segmented. This increase can be also be seen by comparing the per-IoU curves shown in Figure 9 with those in Figure 8, as well as the results presented in Table 2.

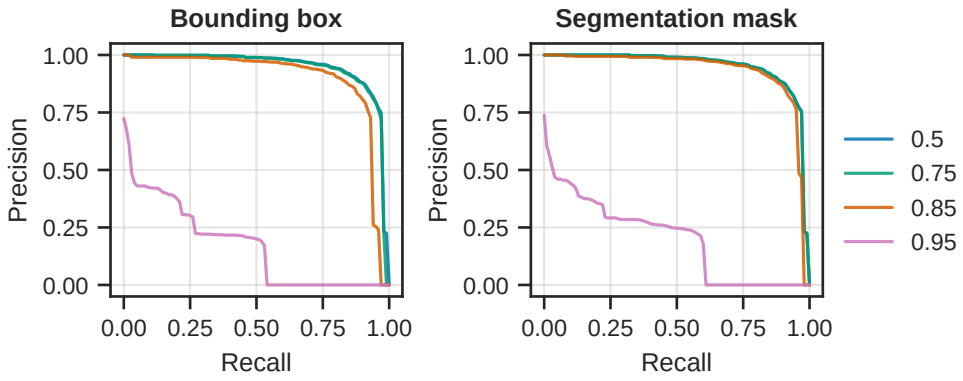


Figure 9. The average PR curves without the "sediment" class, at different IoU thresholds for both tasks. **Left**: Without the "sediment" class, the curves for thresholds $\{0.5, 0.75, 0.85\}$ are almost identical, whereas there is still a major decrease for the 0.95 IoU threshold. **Right**: The PR curves for the segmentation task paint the same picture as for the detection task, indicating that few predictions are correct above 0.95 IoU, and that very many targets are not being predicted.

3.3. Qualitative Analysis of Predictions

When evaluating the predictions manually, it became apparent that the overall accuracy and quality of the segmentation masks produced by the model are good. The boundary of the masks quite precisely delineates the foraminifera and sediment grains

[3] Due to the resources required to annotate more than 7000 objects based on their proximity to other objects, with sufficient precision and recall.

from the background. For the most part, the predicted bounding boxes correspond well with the masks. One of the biggest challenges seem to lie in the classification of object labels; there are (for trained observers) many obvious misclassifications. The exact cause is somewhat uncertain, but in many cases the objects are relatively small and feature-less. It is not hard to image how a feature-less planktic foraminifera can be misclassified as benthic, especially if the object is small. Other cases of misclassifications are likely caused by a lack of training examples; many seem like out-of-distribution examples due to the high confidence score. Examples of predictions can be seen in Figure 10 and Figure 11.

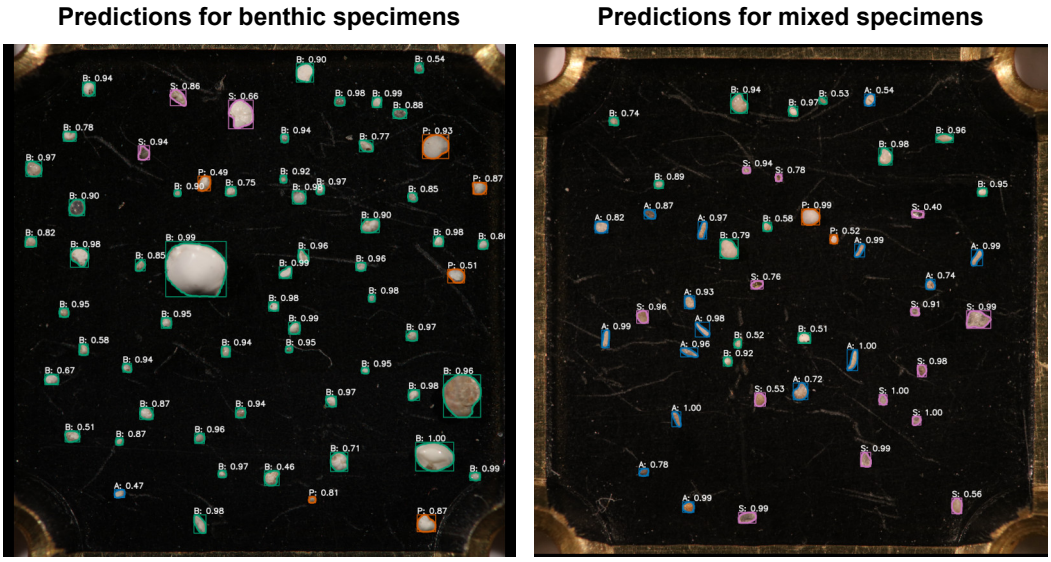


Figure 10. Examples of predictions for two images from the test dataset. **Left**: Predictions for purely calcareous benthic specimens. The accuracy and quality of the predicted masks and bounding boxes are good, but there are several misclassified objects. **Right**: Predictions for a mixture of specimen types. The accuracy and quality of the predicted masks and bounding boxes are good. However, there are misclassified detections for this image as well.

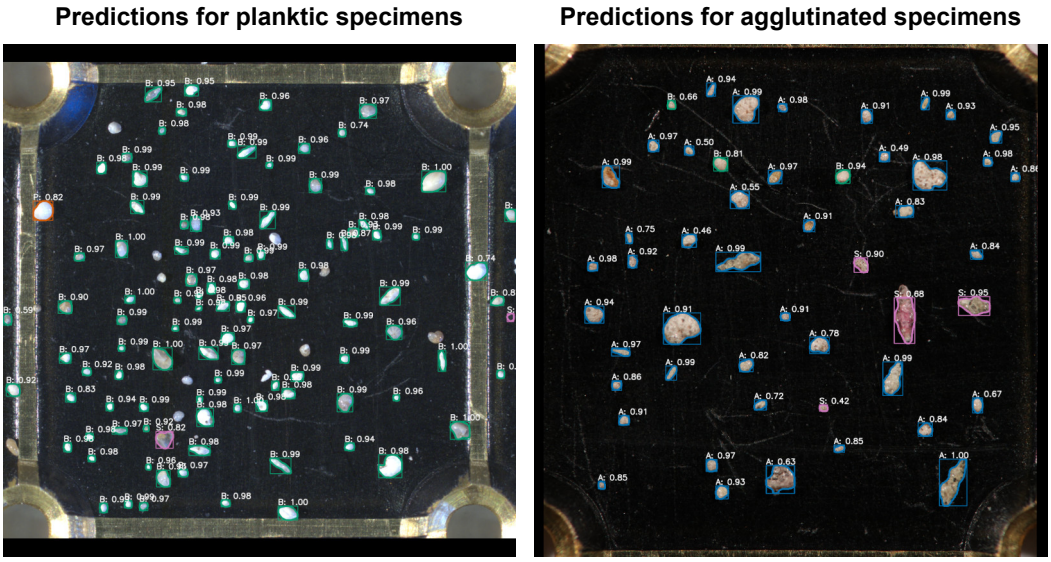


Figure 11. Additional examples of predictions for two images from the test dataset. **Left**: Predictions for purely planktic specimens. There are a few false positives, but the accuracy and quality of true positive detections are good. Note that some objects that have been misclassified. **Right**: Predictions for purely agglutinated benthic specimens. Good accuracy and quality of predicted masks and bounding boxes the majority of detections. However, low quality masks and misclassified detections are visible.

4. Discussion

The results presented clearly show that a model built on the Mask R-CNN architecture is capable of performing instance segmentation of microscopic foraminifera. Using model parameters pretrained on the COCO dataset, we adapted and fine-tuned the model for our novel dataset and achieved AP scores of 0.78 and 0.80 on the bounding box and instance segmentation tasks, respectively. There were significant increases in precision and recall when going from averaging over all IoU thresholds (i.e. AP and AR), to specific IoU thresholds. When evaluated with an IoU of 0.5, precision increased to 0.90 for both tasks, and with an IoU of 0.75 the precision was 0.88 for the detection task and 0.90 for the segmentation task. This means that predicting bounding boxes and segmentation masks that almost perfectly overlap with their respective ground-truth is challenging for the given dataset, and possibly for the model architecture or hyperparameters. It is likely that there are errors in the annotated ground-truth object masks in the dataset, both in terms of inaccuracies at the pixel-level, but also potential false positives or false negatives, meaning that achieving perfect predictions are very unlikely. Importantly, depending upon the specific application of an instance segmentation model, pixel-perfect predictions might not be a necessity.

Omitting the "sediment" class also lead to significant increases in model performance, which can be explained by the challenging nature of some test images that contained very dense clusters of sediment grains. This can in part be mitigated in practical applications by ensuring objects are not clustered, but ideally this also should be addressed at the model-level. It is possible that this can, to some extent, be overcome by introducing much more training examples with crowded scenes, and correctly annotating all objects as being in a crowd. Additionally, it is possible that the issue can also be reduced by tuning the hyperparameters of the Mask R-CNN architecture.

Both quantitative and qualitative analysis of the predicted detections and segmentation masks suggest that the model is performing well. However, the results also show that there are some challenges that should to be investigated further and addressed in future work.

4.1. Future Research

Based on the experiments and results, we propose a few research ideas worth investigating in future efforts.

4.1.1. Expanding and Revisiting the Dataset

Expanding the dataset is perhaps the most natural extension of the presented work. If carefully curated, a more exhaustive dataset should help improve some of the corner cases where the model is struggling to produce accurate predictions. Additionally, with the appropriate resources it would be valuable to ensure every object in the existing dataset is appropriately labeled as part of a "crowd" or not. Improving the accuracy of the e.g. densely packed "sediment" objects, will improve model performance, as well as make the model more applicable to real-world situations. Another important aspect of expanding the dataset is introduce species-level object classes, as opposed to the high-level categories used today. Accurately detecting microscopic foraminiferal species is vital to most downstream geoscience applications.

4.1.2. Additional Hyperparameter Tuning

If sufficient computational resources are available, performing more exhaustive hyperparameter tuning should be pursued. While this should include experiments with optimizers, learning rates, and so forth, it should more crucially be focused on the numerous hyperparameters of the Mask R-CNN model components. Specifically, the parameters of the regional proposal network, and the fully-convolutional network (for mask prediction) should be validated and experimented with. It is entirely possible some number of these parameters are sub-optimal for the given dataset.

4.1.3. Improved GPU Training

While training on multiple GPUs might not lead to big improvements in model performance, the increased effective batch size will help stabilize and speed up training. Additionally, given the small size of the most objects relative to the image dimensions, training without having to resize the images to fit in GPU memory will increase model performance. This could be solved directly by using GPUs with more memory, or possibly by partitioning each image across multiple GPUs, predicting on a sub-region per GPU.

4.1.4. Other Segmentation Models

We chose to use Mask R-CNN primarily because of its capabilities, but also because proven, pre-trained weights were readily available. Recently, numerous models have been published that surpass Mask R-CNN in several performance metrics, and importantly also seem to have much faster inference times (which is important for real-world applications.) Examples of alternative models that should be tested include PANet [16], TensorMask [17], CenterMask [18], SOLOv2 [19].

4.1.5. Uncertainty Estimation

We have shown that the model is robust to training runs with different random seeds, and the next natural step is to investigate robustness with regards to different training/test data splits, and to estimate the uncertainty of the model predictions. Some work has been published on estimating model predictive uncertainty of Mask R-CNN models [27–29]. However, it should be possible to avoid the need for introducing Monte Carlo dropout sampling [30], which requires making changes to existing models, by leveraging the more recent Monte Carlo batch normalization sampling [31] technique instead.

5. Conclusions

The proposed model achieved an AP of 0.78 ± 0.00 on the bounding box (detection) task and 0.80 ± 0.00 on the segmentation task, based on 10 training runs with different

random seeds. We also evaluated the model without the challenging sediment grain images, and the AP for both tasks increased to 0.84 ± 0.00 and 0.86 ± 0.00, respectively.

When evaluating predictions both qualitatively and quantitatively, we saw the predicted bounding boxes and segmentation masks were good for the majority of test cases. However, there were many cases of incorrect class label predictions; mostly for small objects, or objects that we hypothesize can be considered out-of-distribution.

Based on the presented results, our proposed model for semantic segmentation of microscopic foraminifera, is a step towards automating the process of identifying, counting, and picking of microscopic foraminifera. However, work remains to be done, such as expanding the dataset to improve the model accuracy, experimenting with other architectures, and implementing uncertainty estimation techniques.

Author Contributions: Conceptualization, T.J., S.S., K.M. and F.G.; methodology, T.J.; software, T.J.; validation, T.J., S.S. and K.M.; formal analysis, T.J.; investigation, T.J. and S.S.; resources, T.J. and S.S.; data curation, T.J. and S.S.; writing—original draft preparation, T.J.; writing—review and editing, T.J., S.S., K.M. and F.G.; visualization, T.J.; supervision, K.M. and F.G.; project administration, T.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was funded by UiT The Arctic University of Norway.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Prediction Examples

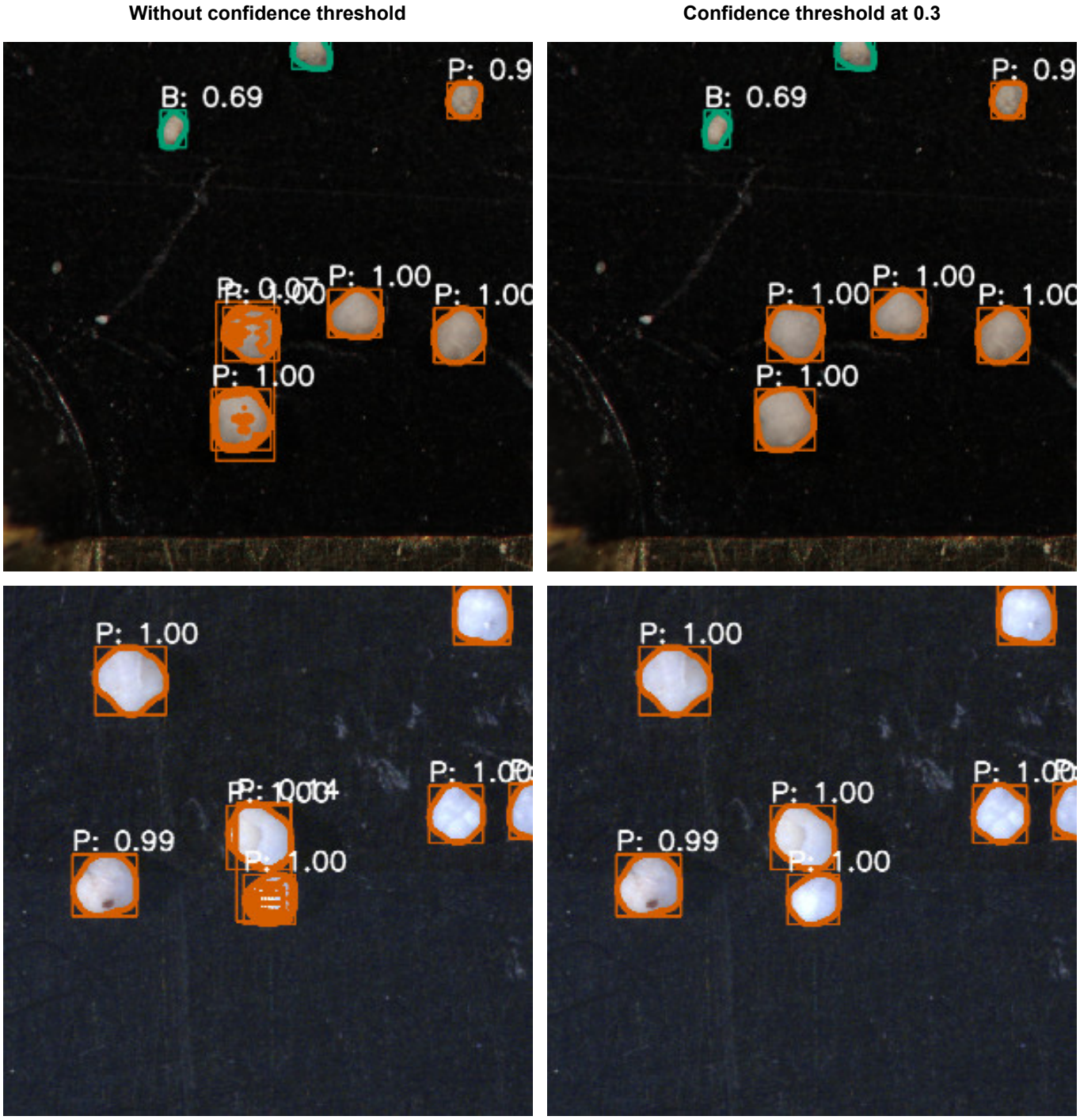


Figure A1. Examples of predicted bounding boxes and segmentation masks for the "planktic" class. **Left column**: Overlapping predictions can be seen near the middle of both images. The confidence score for the overlapping predictions with low-quality masks is significantly lower than the high-quality predictions. We can also see that some smaller objects in the top image have been misclassified as the "benthic" class. **Right column**: The overlapping predictions have been removed by thresholding the confidence score at 0.3.

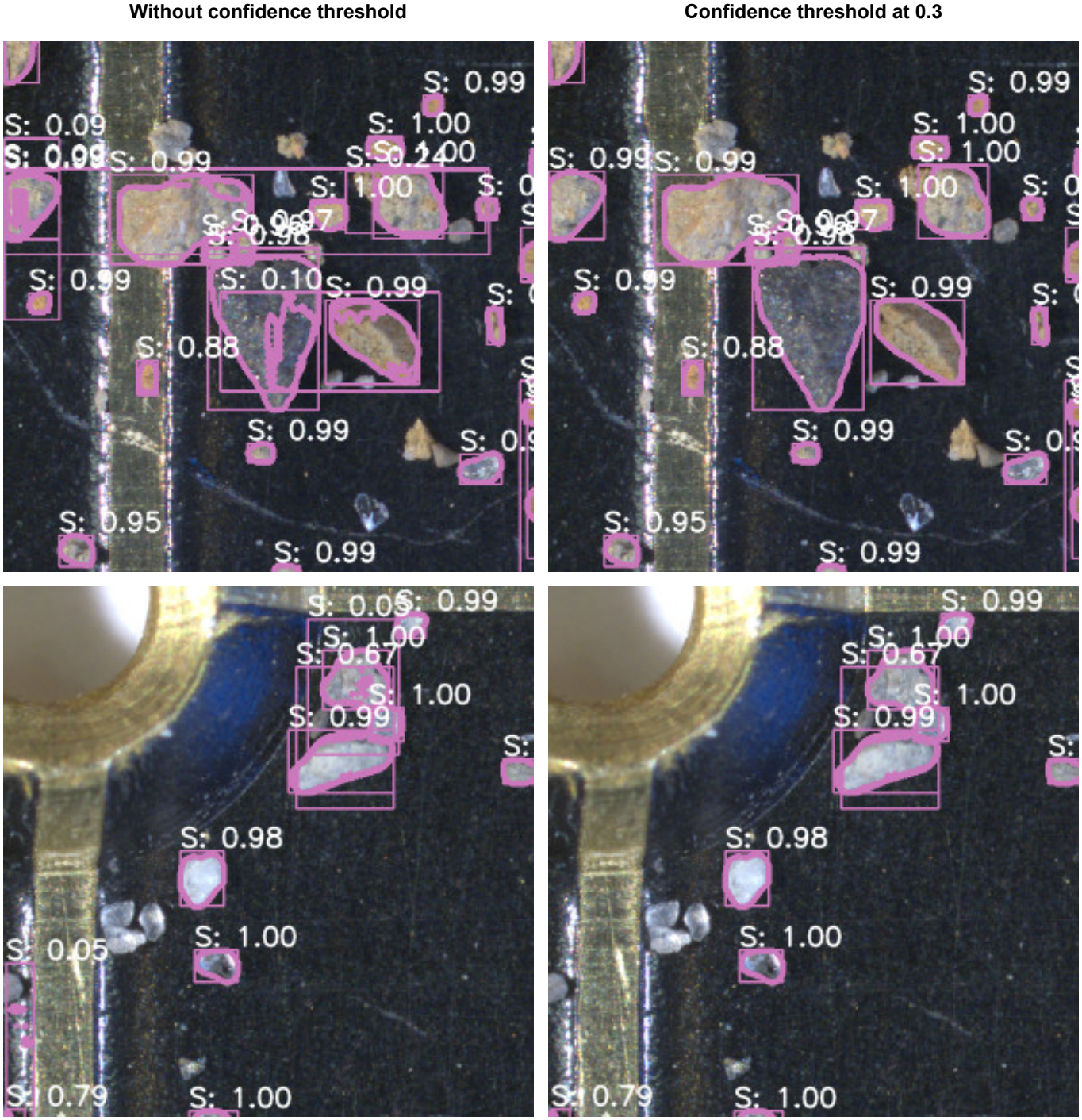


Figure A2. Examples of predicted bounding boxes and segmentation masks for the "sediment" class. **Left column**: Overlapping predictions can be seen near the middle of both images. Also, notice that several objects have been missed entirely. **Right column**: Most of the overlapping predictions have been removed by thresholding the confidence score at 0.3.

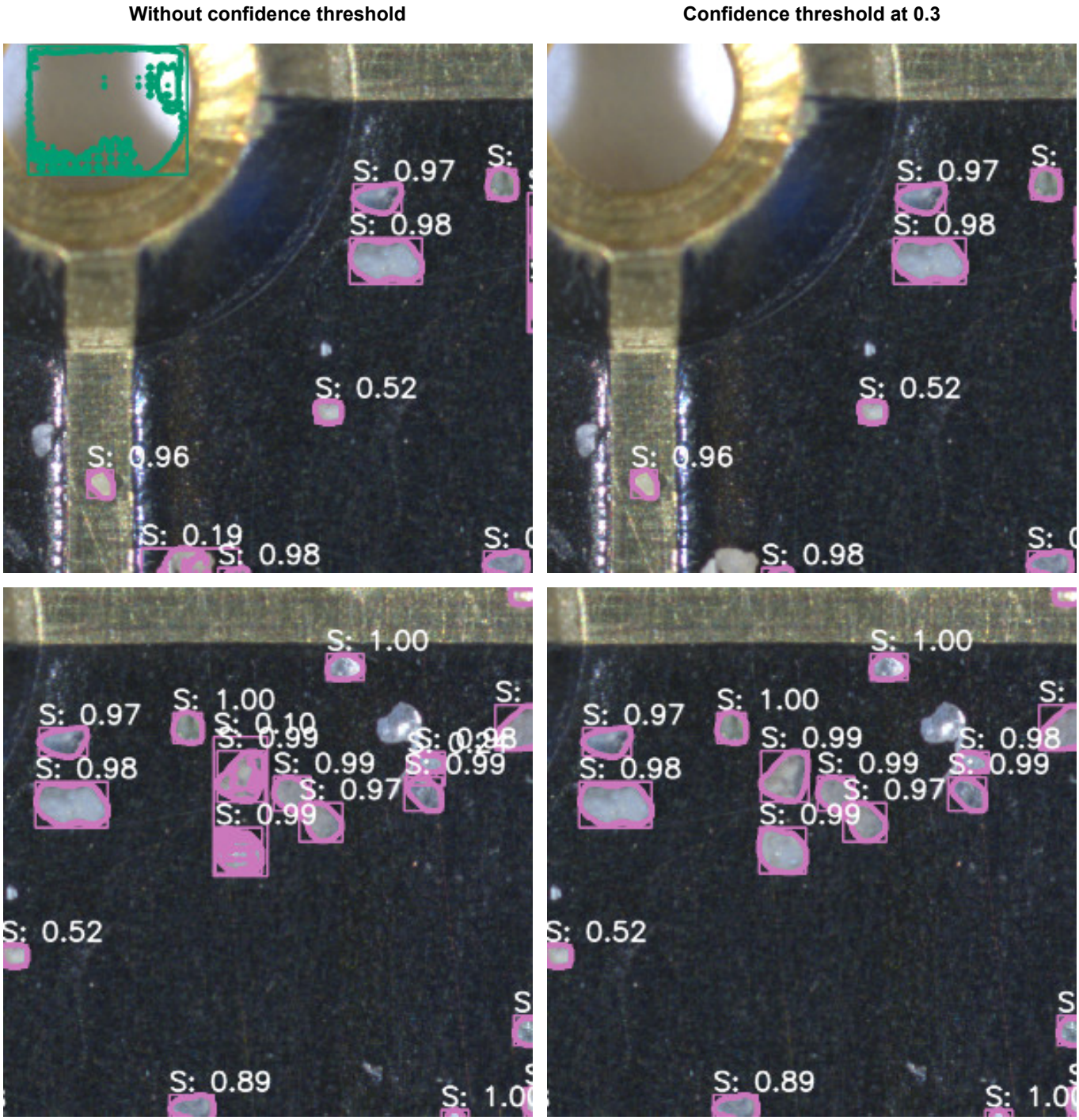


Figure A3. Additional examples of predicted bounding boxes and segmentation masks for the "sediment" class. **Left column**: A very obvious false positive detection of the "benthic" class can be seen near the top-left corner of the first image. For the second image, several overlapping predictions can be seen. **Right column**: The false positive detection and the overlapping predictions have been removed by thresholding the confidence score at 0.3.

References

1. Hayward, Bruce, B.W..; Le Coze, François, F..; Vachard, Daniel, D..; Gross, Onno, O.. World Foraminifera Database. Accessed at http://www.marinespecies.org/foraminifera on 2021-04-08, 2021. doi:10.14284/305.
2. Emiliani, C. Pleistocene temperatures. *The Journal of Geology* **1955**, *63*, 538–578.
3. Hald, M.; Andersson, C.; Ebbesen, H.; Jansen, E.; Klitgaard-Kristensen, D.; Risebrobakken, B.; Salomonsen, G.R.; Sarnthein, M.; Sejrup, H.P.; Telford, R.J. Variations in temperature and extent of Atlantic Water in the northern North Atlantic during the Holocene. *Quaternary Science Reviews* **2007**, *26*, 3423–3440.
4. Katz, M.E.; Cramer, B.S.; Franzese, A.; Hönisch, B.; Miller, K.G.; Rosenthal, Y.; Wright, J.D. Traditional and emerging geochemical proxies in foraminifera. *The Journal of Foraminiferal Research* **2010**, *40*, 165–192.
5. Suokhrie, T.; Saraswat, R.; Nigam, R.; Kathal, P.; Talib, A. Foraminifera as bio-indicators of pollution: a review of research over the last decade. *Micropaleontology and its Applications. Scientific Publishers (India)* **2017**, pp. 265–284.
6. Zhong, B.; Ge, Q.; Kanakiya, B.; Marchitto, R.M.T.; Lobaton, E. A comparative study of image classification algorithms for Foraminifera identification. 2017 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2017, pp. 1–8.
7. Ge, Q.; Zhong, B.; Kanakiya, B.; Mitra, R.; Marchitto, T.; Lobaton, E. Coarse-to-fine foraminifera image segmentation through 3D and deep features. 2017 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2017, pp. 1–8.
8. de Garidel-Thoron, T.; Marchant, R.; Soto, E.; Gally, Y.; Beaufort, L.; Bolton, C.T.; Bouslama, M.; Licari, L.; Mazur, J.C.; Brutti, J.M.; others. Automatic Picking of Foraminifera: Design of the Foraminifera Image Recognition and Sorting Tool (FIRST) Prototype and Results of the Image Classification Scheme. AGU Fall Meeting Abstracts, 2017.
9. Mitra, R.; Marchitto, T.; Ge, Q.; Zhong, B.; Kanakiya, B.; Cook, M.; Fehrenbacher, J.; Ortiz, J.; Tripati, A.; Lobaton, E. Automated species-level identification of planktic foraminifera using convolutional neural networks, with comparison to human performance. *Marine Micropaleontology* **2019**, *147*, 16–24.
10. Johansen, T.H.; Sørensen, S.A. Towards detection and classification of microscopic foraminifera using transfer learning. Proceedings of the Northern Lights Deep Learning Workshop, 2020.
11. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2017**, *42*, 386–397, [1703.06870]. doi:10.1109/TPAMI.2018.2844175.
12. Dutta, A.; Gupta, A.; Zissermann, A. VGG Image Annotator (VIA). http://www.robots.ox.ac.uk/ vgg/software/via/, 2016. Version: 2.0.9, Accessed: 2020-03-10.
13. Dutta, A.; Zisserman, A. The VIA Annotation Software for Images, Audio and Video. Proceedings of the 27th ACM International Conference on Multimedia; ACM: New York, NY, USA, 2019; MM '19. doi:10.1145/3343031.3350535.
14. Girshick, R. Fast R-CNN. 2015 IEEE International Conference on Computer Vision (ICCV). IEEE, 2015, pp. 1440–1448, [1504.08083]. doi:10.1109/ICCV.2015.169.
15. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2017**, *39*, 1137–1149, [1506.01497]. doi:10.1109/TPAMI.2016.2577031.
16. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018, pp. 8759–8768, [1803.01534]. doi:10.1109/CVPR.2018.00913.
17. Chen, X.; Girshick, R.; He, K.; Dollar, P. TensorMask: A Foundation for Dense Object Segmentation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019, Vol. 2019-Octob, pp. 2061–2069, [1903.12174]. doi:10.1109/ICCV.2019.00215.
18. Wang, Y.; Xu, Z.; Shen, H.; Cheng, B.; Yang, L. CenterMask: single shot instance segmentation with point representation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **2020**, pp. 12190–12199, [2004.04446].
19. Wang, X.; Zhang, R.; Kong, T.; Li, L.; Shen, C. SOLOv2: Dynamic and Fast Instance Segmentation. NeurIPS, 2020, pp. 1–17, [2003.10152].
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **2015**, *2016-Decem*, 770–778, [1512.03385]. doi:10.1109/CVPR.2016.90.
21. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017, pp. 936–944, [1612.03144]. doi:10.1109/CVPR.2017.106.
22. Lin, T.Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. Microsoft COCO: Common Objects in Context, 2015, [arXiv:cs.CV/1405.0312].
23. Sutskever, I.; Martens, J.; Dahl, G.; Hinton, G. On the importance of initialization and momentum in deep learning. Proceedings of the 30th International Conference on Machine Learning; McAllester, S.D.; David., Eds. PMLR, 2013, Vol. 28, *Proceedings of Machine Learning Research*, pp. 1139–1147.
24. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. International Conference on Learning Representations, 2015, [1412.6980].
25. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. International Conference on Learning Representations, 2019, [1711.05101].
26. COCO: Common Object in Context – Detection Evaluation. https://cocodataset.org/#detection-eval. Accessed: 2021-04-17.
27. Miller, D.; Nicholson, L.; Dayoub, F.; Sünderhauf, N. Dropout Sampling for Robust Object Detection in Open-Set Conditions. 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018, pp. 3243–3249. doi:10.1109/ICRA.2018.8460700.

28. Miller, D.; Dayoub, F.; Milford, M.; Sünderhauf, N. Evaluating Merging Strategies for Sampling-based Uncertainty Techniques in Object Detection. 2019 International Conference on Robotics and Automation (ICRA), 2019, pp. 2348–2354. doi:10.1109/ICRA.2019.8793821.
29. Morrison, D.; Milan, A.; Antonakos, N. Uncertainty-aware Instance Segmentation using Dropout Sampling. The Robotic Vision Probabilistic Object Detection Challenge (CVPR 2019 Workshop), 2019.
30. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. Proceeding of the 33rd International Conference on Machine Learning. PMLR, 2016, Vol. 48, pp. 1050–1059.
31. Teye, M.; Azizpour, H.; Smith, K. Bayesian Uncertainty Estimation for Batch Normalized Deep Networks. Proc. 35th Int. Conf. Mach. Learn., 2018, Vol. 11, pp. 7824–7833, [1802.06455].

Bibliography

[1] Terrence J. Sejnowski. *The Deep Learning Revolution*. The MIT Press. MIT Press, 2018. ISBN: 9780262038034.

[2] Hyuna Sung et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." In: *CA: A Cancer Journal for Clinicians* 68.6 (Feb. 2021), caac.21660. ISSN: 0007-9235. DOI: 10.3322/caac.21660.

[3] American Cancer Society. *American Cancer Society. Cancer Facts & Figures 2020*. Atlanta, 2020. URL: http://www.cancer.org/acs/groups/content/@nho/documents/document/caff2007pwsecuredpdf.pdf.

[4] B.W. Hayward et al. *World Foraminifera Database*. 2021. DOI: 10.14284/305. URL: http://www.marinespecies.org/foraminifera.

[5] Cesare Emiliani. "Pleistocene Temperatures." In: *The Journal of Geology* 63.6 (1955), pp. 538–578.

[6] Morten Hald et al. "Variations in temperature and extent of Atlantic Water in the northern North Atlantic during the Holocene." In: *Quaternary Science Reviews* 26.25-28 (Dec. 2007), pp. 3423–3440. ISSN: 02773791. DOI: 10.1016/j.quascirev.2007.10.005.

[7] Miriam E. Katz et al. "Traditional and Emerging Geochemical Proxies in Foraminifera." In: *The Journal of Foraminiferal Research* 40.2 (Apr. 2010), pp. 165–192. ISSN: 0096-1191. DOI: 10.2113/gsjfr.40.2.165.

[8] Thejasino Suokhrie et al. "Foraminifera as Bio-Indicators of Pollution: A Review of Research over the Last Decade." In: *Micropaleontology and its Applications*. Ed. by P.K. Kathal, Rajiv Nigam, and Abu Talib. Scientific Publishers, 2017, pp. 265–284. ISBN: 9789387869769.

[9] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks." In: *Nature* 542.7639 (2017), pp. 115–118. ISSN: 0028-0836. DOI: 10.1038/nature21056.

[10] Mihaela Antonina Calin et al. “Hyperspectral Imaging in the Medical Field: Present and Future.” In: *Applied Spectroscopy Reviews* 49.6 (Aug. 2014), pp. 435–447. ISSN: 0570-4928. DOI: 10.1080/05704928.2013.838678.

[11] Guolan Lu and Baowei Fei. “Medical hyperspectral imaging: a review.” In: *Journal of Biomedical Optics* 19.1 (Jan. 2014), p. 010901. ISSN: 1083-3668. DOI: 10.1117/1.JBO.19.1.010901.

[12] Laura Rey-Barroso et al. “Visible and Extended Near-Infrared Multispectral Imaging for Skin Cancer Diagnosis.” In: *Sensors* 18.5 (May 2018), p. 1441. ISSN: 1424-8220. DOI: 10.3390/s18051441.

[13] Qian Ge et al. “Coarse-to-fine foraminifera image segmentation through 3D and deep features.” In: *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. Vol. 2018-Janua. IEEE, Nov. 2017, pp. 1–8. ISBN: 978-1-5386-2726-6. DOI: 10.1109/SSCI.2017.8280982.

[14] R. Mitra et al. “Automated species-level identification of planktic foraminifera using convolutional neural networks, with comparison to human performance.” In: *Marine Micropaleontology* 147.September 2018 (Mar. 2019), pp. 16–24. ISSN: 03778398. DOI: 10.1016/j.marmicro.2019.01.005.

[15] Ross Marchant et al. “Automated analysis of foraminifera fossil records by image classification using a convolutional neural network.” In: *Journal of Micropalaeontology* 39.2 (Oct. 2020), pp. 183–202. ISSN: 2041-4978. DOI: 10.5194/jm-39-183-2020.

[16] Takuya Itaki et al. “Innovative microfossil (radiolarian) analysis using a system for automated image collection and AI-based classification of species.” In: *Scientific Reports* 10.1 (Dec. 2020), p. 21136. ISSN: 2045-2322. DOI: 10.1038/s41598-020-77812-6.

[17] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition.” In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (Sept. 2014), pp. 1–14.

[18] Jia Deng et al. “ImageNet: A large-scale hierarchical image database.” In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2009, pp. 248–255. ISBN: 978-1-4244-3992-8. DOI: 10.1109/CVPRW.2009.5206848.

[19] Kaiming He et al. “Mask R-CNN.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.2 (2018), pp. 386–397. ISSN: 19393539. DOI: 10.1109/TPAMI.2018.2844175.

[20] Tsung-Yi Lin et al. “Microsoft COCO: Common Objects in Context.” In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 8693 LNCS. PART 5. 2014, pp. 740–755. DOI: 10.1007/978-3-319-10602-1{_}48.

[21] Alexander F. H. Goetz et al. “Imaging Spectrometry for Earth Remote Sensing.” In: *Science* 228.4704 (June 1985), pp. 1147–1153. ISSN: 0036-8075. DOI: 10.1126/science.228.4704.1147.

[22] D. H. Sliney. “What is light? The visible spectrum and beyond.” In: *Eye* 30.2 (Feb. 2016), pp. 222–229. ISSN: 0950-222X. DOI: 10.1038/eye.2015.252.

[23] Kurt Nassau. *Colour*. 2020. URL: https://www.britannica.com/science/color.

[24] Ashkan Ojaghi et al. “Ultraviolet Hyperspectral Interferometric Microscopy.” In: *Scientific Reports* 8.1 (Dec. 2018), p. 9913. ISSN: 2045-2322. DOI: 10.1038/s41598-018-28208-0.

[25] J.M.P. Nascimento and J.M.B. Dias. “Vertex component analysis: a fast algorithm to unmix hyperspectral data.” In: *IEEE Transactions on Geoscience and Remote Sensing* 43.4 (Apr. 2005), pp. 898–910. ISSN: 0196-2892. DOI: 10.1109/TGRS.2005.844293.

[26] Martin Hedegaard et al. “Spectral unmixing and clustering algorithms for assessment of single cells by Raman microscopic imaging.” In: *Theoretical Chemistry Accounts* 130.4-6 (Dec. 2011), pp. 1249–1260. ISSN: 1432-881X. DOI: 10.1007/s00214-011-0957-1.

[27] Jun Li, José M. Bioucas-Dias, and Antonio Plaza. “Spectral–Spatial Hyperspectral Image Segmentation Using Subspace Multinomial Logistic Regression and Markov Random Fields.” In: *IEEE Transactions on Geoscience and Remote Sensing* 50.3 (Mar. 2012), pp. 809–823. ISSN: 0196-2892. DOI: 10.1109/TGRS.2011.2162649.

[28] Hannes Kazianka, Raimund Leitner, and Jürgen Pilz. “Segmentation and Classification of Hyper-Spectral Skin Data.” In: *Data Analysis, Machine Learning and Applications*. Ed. by Christine Preisach et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 245–252. ISBN: 978-3-540-78246-9. DOI: 10.1007/978-3-540-78246-9{_}29.

[29] Nahum Gat. “Imaging spectroscopy using tunable filters: a review.” In: *Proc. SPIE 4056, Wavelet Applications VII*. Ed. by Harold H. Szu et al. Apr. 2000, pp. 50–64. DOI: 10.1117/12.381686.

[30] Sascha Grusche. “Basic slit spectroscope reveals three-dimensional scenes through diagonal slices of hyperspectral cubes.” In: *Applied Optics* 53.20 (July 2014), p. 4594. ISSN: 1559-128X. DOI: 10.1364/AO.53.004594.

[31] Nathan Hagen and Michael W. Kudenov. “Review of snapshot spectral imaging technologies.” In: *Optical Engineering* 52.9 (Sept. 2013), p. 090901. ISSN: 0091-3286. DOI: 10.1117/1.OE.52.9.090901.

[32] José Manuel Amigo, Hamid Babamoradi, and Saioa Elcoroaristizabal. *Hyperspectral image analysis. A tutorial*. 2015.

[33] Dorra Nouri, Yves Lucas, and Sylvie Treuillet. “Calibration and test of a hyperspectral imaging prototype for intra-operative surgical assistance.” In: *Medical Imaging 2013: Digital Pathology*. Ed. by Metin N. Gurcan and Anant Madabhushi. Vol. 8676. Mar. 2013, 86760P. ISBN: 9780819494504. DOI: 10.1117/12.2006620.

[34] Freddie Bray et al. “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.” In: *CA: A Cancer Journal for Clinicians* 00.00 (Sept. 2018), pp. 1–31. ISSN: 00079235. DOI: 10.3322/caac.21492.

[35] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. “Cancer statistics, 2018.” In: *CA: A Cancer Journal for Clinicians* 68.1 (Jan. 2018), pp. 7–30. ISSN: 00079235. DOI: 10.3322/caac.21442.

[36] Wilhelm Stolz et al. *Color Atlas of Dermatoscopy*. Wiley-Blackwell, 2002, p. 248. ISBN: 978-1405100984.

[37] Naheed R. Abbasi et al. “Early Diagnosis of Cutaneous Melanoma.” In: *JAMA* 292.22 (Dec. 2004), p. 2771. ISSN: 0098-7484. DOI: 10.1001/jama.292.22.2771.

[38] F. Rilke. “A Modern View of Histopathological Diagnosis and Classification of Cancer.” In: *Surgical Oncology*. Ed. by Umberto Veronesi et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 1989, pp. 62–83. ISBN: 978-3-642-72646-0. DOI: 10.1007/978-3-642-72646-0{_}7.

[39] P. Salmon et al. “Surgical excision of skin cancer: the importance of training.” In: *British Journal of Dermatology* 162.1 (Jan. 2010), pp. 117–122. ISSN: 00070963. DOI: 10.1111/j.1365-2133.2009.09548.x.

[40] Carolyn R. Rogers-Vizena et al. “Surgical Treatment and Reconstruction of Non-melanoma Facial Skin Cancers.” In: *Plastic and Reconstructive Surgery* 135.5 (May 2015), 895e–908e. ISSN: 0032-1052. DOI: 10.1097/PRS.0000000000001146.

[41] J. Caddick et al. “Psychological outcomes following surgical excision of facial skin cancers.” In: *European Journal of Plastic Surgery* 36.2 (Feb. 2013), pp. 75–82. ISSN: 0930-343X. DOI: 10.1007/s00238-012-0748-5.

[42] Toral S. Vaidya et al. “Appearance-related psychosocial distress following facial skin cancer surgery using the FACE-Q Skin Cancer.” In: *Archives of Dermatological Research* 311.9 (Nov. 2019), pp. 691–696. ISSN: 0340-3696. DOI: 10.1007/s00403-019-01957-2.

[43] Michal Kucera. “Planktonic Foraminifera as Tracers of Past Oceanic Environments.” In: *Developments in Marine Geology*. Vol. 1. 07. 2007. Chap. 6, pp. 213–262. ISBN: 9780444527554. DOI: 10.1016/S1572-5480(07)01011-1.

[44] John R. Haynes. *Foraminifera*. Palgrave Macmillan UK, 1981. ISBN: 9781349053971.

[45] Andrew Booth, Anthea Sutton, and Diana Papaloannou. *Systematic Approaches to a Successful Literature Review*. Second. SAGE Publications Ltd, Aug. 2016. ISBN: 978-1473912465.

[46] Maria J. Grant and Andrew Booth. “A typology of reviews: an analysis of 14 review types and associated methodologies.” In: *Health Information & Libraries Journal* 26.2 (June 2009), pp. 91–108. ISSN: 14711834. DOI: 10.1111/j.1471-1842.2009.00848.x.

[47] M. Petticrew. “Systematic reviews from astronomy to zoology: myths and misconceptions.” In: *BMJ* 322.7278 (Jan. 2001), pp. 98–101. ISSN: 09598138. DOI: 10.1136/bmj.322.7278.98.

[48] Iain Chalmers, Larry V Hedges, and Harris Cooper. “A Brief History of Research Synthesis.” In: *Evaluation & the Health Professions* 25.1 (Mar. 2002), pp. 12–37. ISSN: 0163-2787. DOI: 10.1177/0163278702025001003.

[49] Daniel Zhang et al. *The AI Index 2021 Annual Report*. Tech. rep. Stanford: AI Index Steering Committee, Human-Centered AI Institute, Stanford University, 2021. URL: https://aiindex.stanford.edu/report/.

[50] Catherine Pickering and Jason Byrne. “The benefits of publishing systematic quantitative literature reviews for PhD candidates and other early-career researchers.” In: *Higher Education Research & Development* 33.3 (May 2014), pp. 534–548. ISSN: 0729-4360. DOI: 10.1080/07294360.2013.841651.

[51] Tzu-Tsung Wong. “Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation.” In: *Pattern Recognition* 48.9 (Sept. 2015), pp. 2839–2846. ISSN: 00313203. DOI: 10.1016/j.patcog.2015.03.009.

[52] T. Yokota et al. “Lymph node metastasis as a significant prognostic factor in gastric cancer: a multiple logistic regression analysis.” In: *Scandinavian Journal of Gastroenterology* 39.4 (Jan. 2004), pp. 380–384. ISSN: 0036-5521. DOI: 10.1080/00365520310008629.

[53] Turgay Ayer et al. “Comparison of Logistic Regression and Artificial Neural Network Models in Breast Cancer Risk Estimation.” In: *RadioGraphics* 30.1 (Jan. 2010), pp. 13–22. ISSN: 0271-5333. DOI: 10.1148/rg.301095057.

[54] Madara Tirzïte et al. “Detection of lung cancer with electronic nose and logistic regression analysis.” In: *Journal of Breath Research* 13.1 (Nov. 2018), p. 016006. ISSN: 1752-7163. DOI: 10.1088/1752-7163/aae1b8.

[55] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. ISBN: 978-0262035613. URL: https://www.deeplearningbook.org.

[56] Frank Rosenblatt. “The perceptron: A probabilistic model for information storage and organization in the brain.” In: *Psychological Review* 65.6 (1958), pp. 386–408. ISSN: 1939-1471. DOI: 10.1037/h0042519.

[57] Charlie Murphy, Patrick Gray, and Gordon Stewart. “Verified perceptron convergence theorem.” In: *Proceedings of the 1st ACM SIGPLAN International Workshop on Machine Learning and Programming Languages*. New York, NY, USA: ACM, June 2017, pp. 43–50. ISBN: 9781450350716. DOI: 10.1145/3088525.3088673.

[58] Sergios Theodoridis and Konstantinos Koutroumbas. *Pattern Recognition*. Elsevier, 2009. ISBN: 9781597492720. DOI: 10.1016/B978-1-59749-272-0.X0001-2.

[59] Juergen Schmidhuber. “Deep Learning in Neural Networks: An Overview.” In: *Neural Networks* 61 (Apr. 2014), pp. 85–117. ISSN: 08936080. DOI: 10.1016/j.neunet.2014.09.003.

[60] Ken-Ichi Funahashi. “On the approximate realization of continuous mappings by neural networks.” In: *Neural Networks* 2.3 (Jan. 1989), pp. 183–192. ISSN: 08936080. DOI: 10.1016/0893-6080(89)90003-8.

[61] G. Cybenko. “Approximation by superpositions of a sigmoidal function.” In: *Mathematics of Control, Signals, and Systems* 2.4 (Dec. 1989), pp. 303–314. ISSN: 0932-4194. DOI: 10.1007/BF02551274.

[62] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. “Multilayer feedforward networks are universal approximators.” In: *Neural Networks* 2.5 (Jan. 1989), pp. 359–366. ISSN: 08936080. DOI: 10.1016/0893-6080(89)90020-8.

[63] Věra Kůrková. "Kolmogorov's theorem and multilayer neural networks." In: *Neural Networks* 5.3 (Jan. 1992), pp. 501–506. ISSN: 08936080. DOI: 10.1016/0893-6080(92)90012-8.

[64] Kevin Jarrett et al. "What is the best multi-stage architecture for object recognition?" In: *2009 IEEE 12th International Conference on Computer Vision*. IEEE, Sept. 2009, pp. 2146–2153. ISBN: 978-1-4244-4420-5. DOI: 10.1109/ICCV.2009.5459469.

[65] Vinod Nair and Geoffrey E. Hinton. "Rectified Linear Units Improve Restricted Boltzmann Machines." In: *Proceedings of the 27th International Conference on Machine Learning*. Madison, WI, USA: Omnipress, 2010, pp. 807–814.

[66] David E Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." In: *Nature* 323.6088 (Oct. 1986), pp. 533–536. ISSN: 0028-0836. DOI: 10.1038/323533a0.

[67] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization." In: *International Conference on Learning Representations*. 2015. URL: http://arxiv.org/abs/1412.6980.

[68] Ilya Sutskever et al. "On the importance of initialization and momentum in deep learning." In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by Sanjoy Dasgupta McAllester and David. Vol. 28. Proceedings of Machine Learning Research. PMLR, 2013, pp. 1139–1147.

[69] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet classification with deep convolutional neural networks." In: *Communications of the ACM* 60.6 (May 2017), pp. 84–90. ISSN: 0001-0782. DOI: 10.1145/3065386.

[70] Kaiming He et al. "Deep Residual Learning for Image Recognition." In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016-Decem (Dec. 2015), pp. 770–778. ISSN: 10636919. DOI: 10.1109/CVPR.2016.90.

[71] Mingxing Tan and Quoc V. Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." In: *36th International Conference on Machine Learning, ICML 2019* 2019-June (May 2019), pp. 10691–10700.

[72] Dumitru Erhan, Aaron Courville, and Yoshua Bengio. *Understanding Representations Learned in Deep Architectures*. Tech. rep. 2010. URL: http://www.dumitru.ca/files/publications/invariances_techreport.pdf.

[73] Matthew D. Zeiler and Rob Fergus. "Visualizing and Understanding Convolutional Networks." In: *Computer Vision – ECCV 2014*. Ed. by David Fleet et al. Cham: Springer International Publishing, 2014, pp. 818–833.

[74] Anh Nguyen, Jason Yosinski, and Jeff Clune. “Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned By Each Neuron in Deep Neural Networks.” In: (Feb. 2016).

[75] Geoffrey E. Hinton et al. “Improving neural networks by preventing co-adaptation of feature detectors.” In: (July 2012), pp. 1–18.

[76] Anabia Sohail et al. “A multi-phase deep CNN based mitosis detection framework for breast cancer histopathological images.” In: *Scientific Reports* 11.1 (Dec. 2021), p. 6215. ISSN: 2045-2322. DOI: 10.1038/s41598-021-85652-1.

[77] Changqian Yu et al. “BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation.” In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11217 LNCS (Aug. 2018), pp. 334–349. ISSN: 16113349. DOI: 10.1007/978-3-030-01261-8{_}20.

[78] Chuang Gan et al. “VQS: Linking Segmentations to Questions and Answers for Supervised Attention in VQA and Question-Focused Semantic Segmentation.” In: *Proceedings of the IEEE International Conference on Computer Vision* 2017-Octob (Aug. 2017), pp. 1829–1838. ISSN: 15505499. DOI: 10.1109/ICCV.2017.201.

[79] Francesco Martino et al. “Deep Learning-Based Pixel-Wise Lesion Segmentation on Oral Squamous Cell Carcinoma Images.” In: *Applied Sciences* 10.22 (Nov. 2020), p. 8285. ISSN: 2076-3417. DOI: 10.3390/app10228285.

[80] Shaoqing Ren et al. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (June 2017), pp. 1137–1149. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2016.2577031.

[81] Tsung-Yi Lin et al. “Feature Pyramid Networks for Object Detection.” In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017, pp. 936–944. ISBN: 978-1-5386-0457-1. DOI: 10.1109/CVPR.2017.106.

[82] Ross Girshick. “Fast R-CNN.” In: *2015 IEEE International Conference on Computer Vision (ICCV)*. Vol. 2015 Inter. IEEE, Dec. 2015, pp. 1440–1448. ISBN: 978-1-4673-8391-2. DOI: 10.1109/ICCV.2015.169.

[83] Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully convolutional networks for semantic segmentation.” In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 07-12-June (2015), pp. 3431–3440. ISSN: 10636919. DOI: 10.1109/CVPR.2015.7298965.

[84] Boyu Wang et al. “Transfer Learning via Minimizing the Performance Gap Between Domains.” In: *Advances in Neural Information Processing Systems* NeurIPS (2019), pp. 10645–10655.

[85] Xinyang Chen et al. “Catastrophic Forgetting Meets Negative Transfer: Batch Spectral Shrinkage for Safe Transfer Learning.” In: *NeurIPS* NeurIPS (2019), p. 11.

[86] Yuqing Wang et al. “CenterMask: single shot instance segmentation with point representation.” In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Apr. 2020), pp. 12190–12199.

[87] Xinlong Wang et al. “SOLOv2: Dynamic and Fast Instance Segmentation.” In: NeurIPS (Mar. 2020), pp. 1–17.

[88] Sinno Jialin Pan and Qiang Yang. “A survey on transfer learning.” In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (2010), pp. 1345–1359. DOI: 10.1109/TKDE.2009.191.

[89] Yanchao Yang and Stefano Soatto. “FDA: Fourier Domain Adaptation for Semantic Segmentation.” In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Apr. 2020), pp. 4084–4094. ISSN: 10636919. DOI: 10.1109/CVPR42600.2020.00414.

[90] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. “A survey of transfer learning.” In: *Journal of Big Data* 3.1 (Dec. 2016), p. 9. ISSN: 2196-1115. DOI: 10.1186/s40537-016-0043-6.

[91] Christian Szegedy et al. “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning.” In: *31st AAAI Conference on Artificial Intelligence, AAAI 2017* (Feb. 2016), pp. 4278–4284.

[92] Bo Du et al. “Unsupervised transfer learning for target detection from hyperspectral images.” In: *Neurocomputing* 120 (Nov. 2013), pp. 72–82. ISSN: 09252312. DOI: 10.1016/j.neucom.2012.08.056.

[93] Yuan Yuan, Xiangtao Zheng, and Xiaoqiang Lu. “Hyperspectral Image Superresolution by Transfer Learning.” In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 10.5 (2017). ISSN: 1939-1404. DOI: 10.1109/JSTARS.2017.2655112.

[94] Gencer Sumbul, Ramazan Gokberk Cinbis, and Selim Aksoy. “Fine-Grained Object Recognition and Zero-Shot Learning in Remote Sensing Imagery.” In: *IEEE Transactions on Geoscience and Remote Sensing* 56.2 (Feb. 2018), pp. 770–779. ISSN: 0196-2892. DOI: 10.1109/TGRS.2017.2754648.

[95] Yunhui Guo et al. “SpotTune: Transfer Learning through Adaptive Fine-tuning.” In: *CVPR* (Nov. 2018), pp. 4805–4814. ISSN: 23318422.

[96] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. “Do better imagenet models transfer better?” In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2019-June (2019), pp. 2656–2666. ISSN: 10636919. DOI: 10.1109/CVPR.2019.00277.

[97] Zhizhong Li and Derek Hoiem. “Learning without Forgetting.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.12 (2018), pp. 2935–2947. ISSN: 19393539. DOI: 10.1109/TPAMI.2017.2773081.

[98] Fuzhen Zhuang et al. “A Comprehensive Survey on Transfer Learning.” In: *Proceedings of the IEEE* 109.1 (Jan. 2021), pp. 43–76. ISSN: 0018-9219. DOI: 10.1109/JPROC.2020.3004555.

[99] Alex Kendall and Yarin Gal. “What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?” In: *Advances in Neural Information Processing Systems* (Mar. 2017), pp. 5575–5585. ISSN: 10495258.

[100] Armen Der Kiureghian and Ove Ditlevsen. “Aleatory or epistemic? Does it matter?” In: *Structural Safety* 31.2 (Mar. 2009), pp. 105–112. ISSN: 01674730. DOI: 10.1016/j.strusafe.2008.06.020.

[101] Yarin Gal and Zoubin Ghahramani. “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning.” In: *33rd International Conference on Machine Learning, ICML 2016* 3 (June 2015), pp. 1651–1660.

[102] Sergey Ioffe and Christian Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.” In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, 2015, pp. 448–456.

[103] Mattias Teye, Hossein Azizpour, and Kevin Smith. “Bayesian Uncertainty Estimation for Batch Normalized Deep Networks.” In: *35th International Conference on Machine Learning, ICML 2018* 11 (Feb. 2018), pp. 7824–7833.

[104] Tony Lindeberg. “Scale-space.” In: *Wiley Encyclopedia of Computer Science and Engineering*. Vol. 609. Sep 2008. Hoboken, NJ, USA: John Wiley & Sons, Inc., Sept. 2008, pp. 2495–2504. ISBN: 9780470050118. DOI: 10.1002/9780470050118.ecse609.

[105] Lasse Holmström and Leena Pasanen. “Bayesian Scale Space Analysis of Differences in Images.” In: *Technometrics* 54.1 (Feb. 2012), pp. 16–29. ISSN: 0040-1706. DOI: 10.1080/00401706.2012.648862.

[106] Lasse Holmström and Leena Pasanen. “Statistical Scale Space Methods.” In: *International Statistical Review* 85.1 (Apr. 2017), pp. 1–30. ISSN: 03067734. DOI: 10.1111/insr.12155.

[107] Probal Chaudhuri and J. S. Marron. “SiZer for Exploration of Structures in Curves.” In: *Journal of the American Statistical Association* 94.447 (Sept. 1999), pp. 807–823. ISSN: 0162-1459. DOI: 10.1080/01621459.1999.10474186.

[108] Fred Godtliebsen, J. S. Marron, and Probal Chaudhuri. “Significance in Scale Space for Bivariate Density Estimation.” In: *Journal of Computational and Graphical Statistics* 11.1 (2002), pp. 1–21.

[109] Fred Godtliebsen, J.S. Marron, and Probal Chaudhuri. “Statistical significance of features in digital images.” In: *Image and Vision Computing* 22.13 (Nov. 2004), pp. 1093–1104. ISSN: 02628856. DOI: 10.1016/j.imavis.2004.05.002.

[110] Panu Erästö and Lasse Holmström. “Bayesian multiscale smoothing for making inferences about features in scatterplots.” In: *Journal of Computational and Graphical Statistics* 14.3 (2005), pp. 569–589. ISSN: 10618600. DOI: 10.1198/106186005X59315.

[111] Lasse Holmström. “BSiZer.” In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.5 (Sept. 2010), pp. 526–534. ISSN: 19395108. DOI: 10.1002/wics.115.

[112] Julio M. Duarte-Carvajalino, Miguel Vélez-Reyes, and Paul Castillo. “Scale-space in hyperspectral image analysis.” In: *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XII*. Ed. by Sylvia S. Shen and Paul E. Lewis. Vol. 6233. May. May 2006, p. 623315. ISBN: 0819462896. DOI: 10.1117/12.667964.