Freie Universität Berlin

Institute for Veterinary Pathology

Work title:

“Automated Diagnosis of Seven Major Skin Tumors in Canines Using a ConvolutionalNeural Network (CNN) on H&E-Stained Whole Slide Images (WSI)”

Inaugural dissertation for the degree of

Doctor Medicinae Veterinariae

At the Freie Universität Berlin

Presented by

Marco Antonio Fragoso García

Journal Nr. 4370

Berlin, 2022

Printed with permission of the Department of Veterinary Medicine of the Freie Universität Berlin

Dekan: Univ.-Prof. Dr. Uwe Rösler

Supervisor: Univ.-Prof. Dr. Robert Klopfleisch

First reviewer: Prof. Dr. Vitaly Belik
Second reviewer: PD Dr. Kerstin Müller

Keywords: dogs, skin, neoplasms, melanoma, skin diseases, skin cancer, diagnostic techniques, imagery, image analysis, digital photography, computer analysis, machine learning, immunohistochemistry, veterinary medicine, pathology.

Day of dissertation: 25.11.2022

Contents

Abbreviations

AI. Artificial intelligence

ANN. Artificial neural network

API. Application programming interface

BMP. Windows bitmap

CAD. Computer-aided diagnosis

CAP. Computer-aided pathology

CAPTH. Computational pathology

CCD. Charged coupled device

CM. Confusion matrix

CNN. Convolutional neural network

CPU. Central processing unit

DBN. Deep belief networks

DC. Digital camara

DICOM. Digital Imaging and Communication in Medicine

DL. Deep learning

DM. Digital microscope

DNN. Deep neural networks

DP. Digital pathology

FITS. Flexible Image Transport System

FN. False negative

FP. False positive

GIF. Graphics Interchange Format

H&E. Hematoxylin and eosin

IHC. Immunohistochemistry

JPEG. Joint Photographic Experts Group

LED. Light-emitting diode

MCT. Mast cell tumor

ML. Machine learning

PNG. Portable network graphics

PNST. Peripheral nerve sheet tumor

RAM. Random access memory

RAW. Raw image format

RM. Robotic microscope

ROI. Region of interest

SCC. Squamous cell carcinoma

TIFF. Tagged Image File Format

TFT. Thin-film transistor (monitor)

TN. True negative

TP. True positive

VM. Virtual microscopy

VMS. Virtual microscopy systems

WSI. Whole slide image

1. Introduction

1.1 Digital Pathology

1.1.1 Overview and definitions

Histopathology is a subfield of pathology, which studies the diagnosis of diseases through the visualization and interpretation of tissues in glass slides stained with hematoxylin and eosin (H&E). It is an essential tool for pathological diagnosis since the 19th century (Poynter, 1967; Turk, 1993). Because it is an accessible and useful tool, the review of H&E-stained slides is considered the gold standard for many diseases (Li et al., 2021). Specifically, in oncology, histopathology is the most important diagnostic tool, as through it, the observer (pathologist) can determine a definitive diagnosis (in most cases accurate), evaluate resection margins, identify micro-metastasis and perform tumor staging. However, to reach an even more precise diagnosis, histopathology relies on multiple diagnostic tests, mainly molecular (e.g., IHC). This has caused the conventional use of histopathology to lose its perceived usefulness (Harris & McCormick, 2010).

During the past few decades, the visualization of H&E-stained glass slides has evolved into a novel concept known as digital pathology (DP), which relies on the production of digital images via cameras (photographic and video) or scanners for digital visualization. The initial concept that began to be used in the 1960s was telepathology. Telepathology was defined at that time as a tool for viewing H&E-stained glass slides shared via telecommunication mediums to achieve remote transmission of data in the form of images over long distances. Its advantages were vast as it contributed greatly to education and research focused on pathology (Farahani & Pantanowitz, 2016). The use of telepathology depends primarily on the transformation of biological or pathological information captured by medical experts (pathologists) through imaging devices such as video and photography, as well as in the use of such information for research, diagnostic or educational purposes (Weinstein et al., 1987).

Although the introduction of DP was attributed to telepathology, these two terms differ in some details; for example, DP is distinct from telepathology in that through the captured images, the pathologist can perform various types of analysis, as well as manage and store the data on large servers. In the case of telepathology, this is not possible. In addition to telepathology, DP was born with the advent of virtual microscopy (VM). VM is the technique of digitizing whole or partial glass slides for complete visualization and handling.

Ronald S. Weinstein first used the name telepathology in 1986, and he is now considered the father of telepathology. He was also the first to outline the steps to be followed for the correct remote visualization of digitized glass slides in diagnostic pathology laboratories (Weinstein, 1986); therefore, he owns the U.S. patents for telepathology for diagnostic purposes.

Nordrum and Eide also contributed to the development of DP, as they were the first to establish a sustainable clinical telepathology service in Norway in 1989 (Nordrum et al., 1991), which is still in operation.

Currently, telepathology still has numerous clinical applications, such as remote histopathological diagnosis (Dunn et al., 2009), frozen specimen diagnosis (Evans et al., 2009), consultation (Graham et al., 2009), diagnosis support to subspecialties like dermatology (Massone et al., 2008), preclinical toxicology research studies (Siegel et al., 2018), education (Dee,2009), among others. Telepathology also improves the efficiency of diagnosis, as in shorttime, histopathological images can be shared to highly specialized pathologists (e.g. nephropathologist, neuropathologists or dermatopathologists) for consultations when a rapid diagnosis is required and when the specialist is not physically available, as happensfrequently in the case of frozen section diagnosis.

Already at the beginning of the first decade of the 21st century, videoconferencing systems and virtual microscopy software began to be used. This allowed the interactive visualization of images through a screen (Farahani et al., 2016).Nevertheless, although these technologies proved to be very practical at that moment, theintroduction of these technologies had its disadvantages as well, due to the high cost of high-speed data transmission networks. In Spain, for example, to solve these problems and allow the users to transmit images in real time, the detail and quality of the images was sacrificed by using compression system and by sending the images using conventional software for Microsoft Windows or via webcam (NetMeeting), provoking more source failures and diagnostic errors compared with current systems (Alfaro- Ferreres, 2001).

Telepathology currently has several resources of telecommunication systems that allow real time collection of patient samples over long distances such as static images (photographs), dynamic virtual samples from the digitalization slides (whole slides images, WSI) or images transmitted by a robotically controlled light microscope in real time (Weinstein et al., 2009). Telepathology can also integrate other routine elements in diagnosis including generation of a written report, instant messaging, quality control of light microscopy and inter-consultation through videoconferencing. This last is still in virtual private networks.

Conversely, studies show that diagnostic accuracy rates are higher when using dynamic pathology rather than static pathology using virtual slides. This is because static telepathology relies on images sent by the pathologist, who shares selected parts of the slide for consultation. In contrast, in dynamic telepathology (also known as VM or DP), the medical experts are able to analyze the whole slide with several magnification levels options, imitating the routine use of a light microscope (Kaplan et al., 2002).

Another huge advantage of DP is the construction of databases, as well as virtual libraries that allow the creation of catalogs sorted by, for instance, disease and tissue type. This also has the potential to increase the quality of diagnosis by allowing the pathologist to access all cases for comparison and re-evaluation in a short time if it is required, even when the pathologist is not physically available. In the same way, these virtual libraries allow the archiving of a large number of cases which are mainly used for research and educational purposes (e.g., pathology training). In addition, in order to have a successful DP diagnostic laboratory, it must be ensured that all necessary storage, security and medical data management procedures are performed and available (Kaplan et al., 2002). This can be achieved by the creation of metadata (patient information), which canbe included into the WSI database.

Contemporary currents in the field of training of future pathologist are focused on the digitalization of their pathology classes with the attempt to replace optical microscopes with computers or other devices (e.g. smartphones, tablets, etc.), in addition to the construction of virtual libraries of slides. A clear example of this are the digital pathology laboratories created using cloud computing technologies or the University of Iowa's "Virtual SlideBox" image repository (Dee & Fales-Williams, 2005).

Following the development of telepathology and its everyday use in many institutions and laboratories, the early 1990s saw the emergence of VM in several areas of life science research (Kumar et al., 2004). VM is defined as a method of transforming histological images (whole slides or fragments) into digital information with a resolution similar to that of conventional optical microscopy (Brochhausen et al., 2015). WSI are created with special scanners; however, digital “slides” can also be formed by attaching together several photomicrographs through the microscope (patches). Both techniques allow for the presentation of histological data through computer networks with extraordinary resolution; however, due to their size, the files usually generate problems during the storage and sending of information. VM is also used in other branches of medicine like histology, hematology and biology.

By using computer technology, DP started utilizing VM as its basis. With the practice of WSI, glass slides are converted into digital images that can be observed, managed, shared and analyzed on a computer screen regardless of the location of the viewer. In recent decades, with the advent these new forms of visualization and the growing era of artificial intelligence (AI), the perception of diagnostic histology has evolved considerably, regaining its value and accessibility (Bertram & Klopfleisch, 2017; Pantanowitz et al., 2018; Abels et al., 2019).

In summary, DP is a recent subfield of pathology based on the transformation of glass slides into digital slides (WSI), with the advantages of manage, visualize, analyze and share image data in a practical and objective way. WSI can be visualized and analyzed by pathologists regardless of the location and has been proved that it can considerably increase diagnosis efficiency by the combination with AI.

1.1.2 Hardware (scanners and scanning)

The challenges faced by those aiming to develop software solutions for automated tumor-focused surgical biopsy diagnostics are numerous. From those related to sample processing (e.g. standardization of H&E stains, microtome artifacts, staining reagent variability) or morphological variance/complexity of the tumors, to those related to hardware (scanning, uneven illumination, focusing) and software validation. These lasts ones are of vital importance for the proper development of effective and efficient solutions with a reliable market application. Most scanners in market are standardized; they allow scanning of complete slides in high resolution and magnification in a short time, with some variations depending on the scanner type and vendor (Jahn et al, 2020). In the same way, the accessibility of software for WSI analysis varies according to the needsand budget of the users. Most of these are focused on traditional H&E and IHC image analysis instead of creating DL algorithms for specific tasks.

For the successful and advantageous practice of DP, the laboratory must be assured that the remote pathologist has appropriate access to pertinent diagnostic material. The process of DP imaging consists of many steps. These operations include some basic steps such as sample preparation and staining, performed in a histologic laboratory; H&E staining is the most popular stain to examine tissue sections. Subsequently, the optical image is transformed into digital information by several digital acquisition options, such as photographic cameras, video cameras and automatic microscopic slide scanners. Finally, this information is processed, compressed and transmitted as an image file through telecommunication networks and to be presented on the health professional's screen (Kaplan et al., 2002; McCullough et al, 2004).

VM systems (VMS), commonly understood as those capable of fully digitizing histologic and cytologic slides or WSI are accessible today in multiple formats and commercial solutions. In 1997, the Department of Computer Science at the University of Maryland and the Department of Pathology at Johns Hopkins Hospitals (Baltimore, USA) described the first operational VMS (Afework et al., 1998; Ferreira et al., 1997). Today, it is possible to digitize all types of histological or cytological slides, from thin (5 µm) paraffin-embedded tissue sections to thick (15 µm) IHC or immunofluorescence sections.

Depending on their purpose, digital imaging solutions can be classified into digital microscopes (total preparation scanning) and diagnostic support systems. The purpose of the DM is the creation of digital slides with full viewing capability at high magnification

(Zarella et al., 2019) and the purpose of the diagnostic aid systems is to aid the localization of the area of interest in throughout all the slide, as well as to objectively quantify histological features (e.g. fibrosis). Both systems allow digitizing the entire slide (or a specific region) and photographing specific fields (depending on the objective of the study or type of diagnosis).

Based on device components DP hardware are divided into robotic microscopes (RM) and scanners.

- RM. The RM maintain their original functionality and components including oculars, multiple objectives (motorized revolver), light control, and position and focus control (Collins et al., 2020), but with a camera assembled to the microscope and an image viewing and analysis software (discussed in the software subtopic).
- Scanners. A computer in/outside the scanner box controls slides scanners and scanning. It differs from the RM by the absence of oculars and position or focus control, as well as the addition of an anti-vibration mechanism (Thrall et al, 2015).

Both devices are capable of generating high-resolution images and both are structurally composed of an optical microscopy system (similar to that of a conventional light microscope), a capture system (snapshots and photography) and a visualizer and controller software.

As the camera integrated inside the device (RM or scanner) is a critical factor in the quality and speed of image acquisition, it is essential to know its basic characteristics and functionality. Slide digitizer usually consist of a CCD sensor (charged coupled device), which generates analog signals (similar to digital cameras, DC) and determinates the quality and resolution of the image, i.e. it establishes the number of pixels detected. DCs inside the slide digitizer do not need the utilization of capture cards because they are connected to the PC through a firewire-port (Kim et al., 2020).

Another important element for correct imaging are high precision and high-speed stages. Their job is to allow the achievement of optimal movement speed without losing image quality (Jones-Hall et al., 2021).

In order to correctly visualize the images collected by the scanning systems, DP solutions use high-resolution monitors that are usually flat TFT with different sizes.

One of the most important factors to determinate the digitization efficiency is the scanning time; however, their objective evaluation is very difficult, as the following aspects must be considered:

- Size of the slide or the tissue area to be scanned.
- Magnification (20x or 40x).
- Size of the CCD.
- Time of the focusing and preview phase.
- Number of focus points
- Data capture speed (from camera to computer and from computer to storage), among others.

The focus map, which counts the number of focused points, can be assigned automatically and manually on most scanners. In Aperio ScanScope it is possible to manually add focus points to those that are automatically detected by the visualizing software.

In some cases (e.g. cytology slides or thick histology), it is necessary to digitize multiple planes of focus, known as the Z-axis. This allows the pathologist to visualize several planes from top to bottom, similar to the use of the micrometer of the conventional microscope.

Before the scanning, a previsualization of the slide is necessary. In this step, it can be decided whether the digitization will be performed on the entire slide or on a specific area (manually) or if only the areas with relevant material will be automatically selected. Most of the scanners avoid blank areas (empty spaces) during this step. Previsualization is usually done in matter of seconds.

Once the number of points to be scanned is determined, the slide is scanned. Digitization consists of capturing areas of the slide and then stitching the fragments together to create the virtual image, which can be adjusted manually or automatically (default settings). The software of each scanner usually supports the scanning process; it is performed from the upper left corner to the lower edge of the slide, in order to create an image with multiple quadrants, like a mosaic. In the case of Aperio ScanScope, the scanning process is linear. By navigating throughout the WSI, the observers are able to perform similar movements as on a microscope, such as lateral and vertical (X and Y axis), as well as to modify the magnification and focus or change the plane of focus (Z axis).

Note: There are no specific publications that analyze and compare all types of scanners with an objective method, so the reader is advised to look into the individual web sites and manuals of each scanner company in order to obtain more information.

1.1.3 Software

Once the slide is digitized, the file is created and can be viewed with different types of specialized software, which usually compresses the images to an optimal size for proper high-resolution viewing on a screen. Regarding the software, the viewer can also perform specific tasks on it for digital pathology and image analysis for diagnosis and investigation purposes. After that, the WSI must be archived but not as a physical slide, but into a dedicated virtual space such as internet servers or inside the computer, depending on the vendor. Both options are useful but it depends on the needs of the user and the accessibility to internet connection and storage capacity.

WSI analysis software is developed with stand-alone and functional AI programs; however, some of them require the download of plugins (complementary programs that extend the functions of web applications and desktop programs) to improve their functionality.

The following is a list of the main free open sourcesoftware most commonly used in DP.

- QuPath. QuPath was first designed to image analysis of WSI (mainly for biomarkers and IHC), but it is currently used also for neoplasia analysis on H&E. This software is easy to use and it includes the possibility to create smart annotations through different types of tools that transform coordinates and pixels into data in order to develop algorithms and analyze the whole tissue. It has also a very useful tool that allows the user to create even more precise annotations in the slide, making this process more efficient (e.g. automatic delineation tissue types and delineation of structures). QuPath offers the users several type of algorithms for solving simple tasks like automatic cell detection or stain estimation (useful for IHC/fluorescence) and supports developers to implement new applications by exchanging data with other software options such as ImageJ (Bankhead et al., 2017).

- Image J/Fiji. ImageJ was first designed with an open architecture that provides extensibility via Java plugins and scriptable macros. This provides the users the possibility to resolve image problems related to processing and analysis by comparing multiple system data. It also has an automated hematology system. This software can display, edit, analyze, process, save, and print 8-bit (256 colors), 16-bit (thousands of colors), and 32-bit (millions of colors) images. It can read various image formats including TIFF, PNG, GIF, JPEG, RAW among others. It is possible to perform tasks on several images in a single window (limited by available memory) and even in various parallel CPUs at the same time. User often use ImageJ to calculate areas and pixels of tissue structures with statistical purposes by measuring exact distances and angles throughout the whole slide and by creating histograms and profile line plots. ImageJ has tools to manipulate contrast, detect edges, perform Fourier analysis, as well as geometric transformations of the slide, such as rotation and flips (Rueden et al., 2017; Schindelin et al., 2012; Schneider et al., 2012).

- CellProfiler. CellProfiler was first designed to aim biologist to analyze and quantify phenotypes from images in an automatic manner and can read most of the WSI formats. In this software, there are several types of algorithms for image analysis like automatic identification, segmentation and measurement of biological structures, which can be used individually or sequentially in a pipeline. CellProfiler can collaborate with some scientific libraries for mathematical operations purposes (Lamprechtet al., 2007).

- Ilastik. Ilastik was designed for image classification and segmentation through the annotation of histological structures and the creation of an automatic classifier. Ilastik has module for using classifiers to process images within a CellProfiler framework (Sommer et al., 2011).

- Orbit. Orbit was first designed for the quantification of large images through analysis algorithms using ML, as well as segmentation and classification of histological structures. In addition, a versatile API allows the owners to enhance Orbit and run their own scripts. Another interesting function of this program is the calculation of different tissue classes' proportion, e.g. the percentage of collagen in a tissue. ML-based tissue quantification allows the pathologist to train the systemon specific tissue classes and quantify them, likewise, segmenting, overlapping and calculating objects features in order to achieve a more accurate classification (Goldberg et al., 2005).

- Cytomine. Cytomine was first designed as a web-based tool for large-scale image-based studies in multidisciplinary teams. It has annotation, analysis and management possibilities in WSI. It can be used in most of the formats by the conversion of the images during the loading stage, which are archived into the cloud, so it provides the option to organize, analyze, explore and share WSI over the internet for collaborative projects. Cytomine includes algorithms (default), but the user can develop other types. Another tools allow the visualization of several images and annotations at the same time, the management of annotations (reviewing, searching, filtering, sorting) even from different creators, creation of DL algorithms, size calculation of structures, among others (Marée et al., 2016).

- Icy. Icy (colloquially defined as the image analysis "photoshop") was designed to visualize, analyze, annotate and quantify features in WSI and other typed bio-images. In this software, researchers and users can also develop algorithms according to their needs (De Chaumont et al., 2012).

Note: There are no specific publications that analyze and compare all types of WSI softwares with an objective method, so the reader is advised to look into the individual web sites and specific citations of each software company in order to obtain more information.

1.2 Artificial intelligence

1.2.1 Overview and definitions

Artificial intelligence (AI) is a computer science field that is defined as the intelligence expressed by machines through their processors and different types of software. These in turn perform functions and tasks that would be the equivalents of the human body, brain and mind in order to behave in a natural way as humans and certain types of animals with complex brains would (Kaplan et al., 2021).

The origin of AI is considered to date back to man's attempts since ancient times to enhance his physical and intellectual potential through the creation of devices with automatisms, emulating the form and abilities of human beings. In computer science, an ideal "intelligent" machine would be one with flexible abilities that perceive its environment and, in turn, carry out actions that maximize the chances of success in some purpose or task (Lopez-Rubio et al., 2015). Andreas Kaplan and Michael Haenlein define artificial intelligence as "the ability of a system to correctly interpret external data, to learn from that data, and to use that knowledge to achieve specific tasks and goals through flexible adaptation" (A. Kaplan & Haenlein, 2019). However, colloquially the term artificial intelligence is used in cases where a machine successfully reproduces certain cognitive functions that humans relate to other human minds, such as perception, reasoning, learning and problem solving (Russell & Norvig, 2002).

Over time, the definition of AI has evolved, as what was once thought to be a task requiring a certain level of intelligence is now considered a common task, such as optical character recognition, tasks that are now commonly used in different branches of technology. Currently, with the development of expert computer systems, the management and control of robots and processors, AI has become a novel way to address problems through the integration and analysis of knowledge shared by the human mind, with certain autonomy to such an extent that machines develop an intelligent system capable of developing its own program. An expert system is defined as a programming structure with the capacity to store and use the complete knowledge about a certain field of study, as well as its translation into computer language and its automatic learning (Patterson, 1990).

Likewise, with the evolution of new technologies and mathematical calculations, AI is also ultimately defined as the ability of machines to use algorithms, learn from data. This way of behavior should mimic to that of a human being. One of the main focuses of artificial intelligence is machine learning (ML), in such a way that computers or machines have the ability to learn without being programmed to do so.

In 1956, John McCarthy first coined the expression "artificial intelligence", and defined it as "the science and ingenuity of making intelligent machines, especially intelligent computer programs" (McCarthy, 2007). Other definitions and points of view are, for example, according to Takeyas (Takeyas, 2007), AI is a branch of computational sciences in charge of studying computational models capable of performing human activities based on two of their primary characteristics: reasoning and behavior. This current definition involves not only the way machines reason, but also their ability to perform tasks that resemble human behavior, as in the case of robotics. There are also other types of perceptions that can be obtained and produced, respectively, by physical sensors and mechanical sensors in machines, electrical or optical pulses in computers, as well as by bit inputs and outputs of software and its software environment.

Several examples are found in the area of system control, automatic planning, the ability to respond to diagnostics and consumer queries, handwriting recognition, speech recognition and pattern recognition. AI systems are now part of the routine in fields such as economics, medicine, engineering, transportation, communications, and the military, and have been used in a variety of computer programs, strategy games such as computer chess, and other video games.

Stuart J. Russell and Peter Norvig diversify several types of artificial intelligence (Russell & Norvig, 2002):

- Systems that think like humans or, more precisely, systems that try to emulate human thinking, as in the case of artificial neural networks. The automation of activities that we link to human thought processes includes, for example, decision making, problem solving and learning (Krogh, 2008).
- Systems that attempt to act like humans or mimic human behavior. The most commonly known example is robotics, defined as the branch of AI that studies how machines manage to perform tasks that currently human beings do better, like in in the field of medicine (Dario et al., 1994).
- Systems that think logically and that seek to mimic the rational thinking of human beings; for example, expert systems, the study of the computations that make it possible to perceive, reason and act (Horvitz et al., 1988).
- Systems that act rationally and that try to emulate rationally human behavior; for example, intelligent agents, which is involved with intelligent behaviors in artifacts (Poole & Mackworth, 2010).

AI is divided into two lines of reasoning:

On the one hand, conventional or symbolic and deductive AI that is defined as the proper and statistical analysis of human behavior for solving different types of problems (Garnelo & Shanahan, 2019). In this type of AI there is (Confalonieri et al., 2021):

- case-based reasoning to support decision making when solving specific problems,
- expert systems that conclude a solution through prior knowledge regarding the context in which it is applied,
- bayesian networks that suggest procedures through probabilistic inference,
- behavior-based AI that is characterized by being autonomous with the ability to self-regulate and control itself to achieve significant improvements when solving tasks
- and smart process management, which provide intelligent support during complex decision making.

These types of solutions suggest solutions to multiple types of problems, equivalent to specialists in the field of interest (Confalonieri et al., 2021).

On the other hand, computational AI or subsymbolic-inductive AI involves development or interactive learning. This type of thinking and learning is based on empirical data. This type of computational intelligence has two purposes, its scientific goal is to glimpse the principles that enable intelligent behavior in both natural and artificial systems and its technological goal is based on the specification of methods to design intelligent systems. This type of technology is mostly used in environmental sciences, climatology and financial markets (Siddique & Adeli, 2013).

Since AI is being used in many technological fields and especially in medicine, its utilization and development is strictly regulated by laws that establish rules and norms of behavior and usability to ensure social welfare and protect individual rights (Keskinbora, 2019). As in any other scientific field, this is done with the aim of minimizing risks and promoting the benefit to society. Although there are currently no legal norms that truly regulate AI, in April 2021, the European Commission externalized a proposal for its regulation in the European Union (Stöger et al., 2021).

The technologies that have been born through AI today are numerous and are found in almost all fields of research, because when a problem is solved by AI, the solution is routinely incorporated into biological and industrial fields (Becker, 2019). In this document we will focus on the advantages of AI in the medical fields, especially in the pattern recognition of digital pathological images.

1.2.2 Machine learning

Machine learning (ML) a computer science subfields and a branch of AI that studies the methods of learning through the use of data (Nichols et al., 2019). Themain goal of ML is to encourage and develop learning models that have the ability to generate results that can in turn improve upon their own experience; in other words, whenthe skill was not present prior to training. In ML, a computer observes and analyzes datawith the aim to develop a model that is able to hypothesize and design software for problem solving. ML is also broadly related to pattern recognition to emulate the scientific method with mathematical techniques.

Applications of ML are currently numerous and include search engines, medical diagnostics, financial fraud detection, stock market analysis, DNA sequence classification, speech and written language recognition, video games and robotics (Kononenko, 2001; Lee et al., 2018).

There are several automated learning models (geometric, probabilistic and logical), some of these seek to eliminate the exhaustive need for expert knowledge in data analysis methods, while others are concerned with the establishment of a collaborative framework between the expert and the computer, as commonly observed in biomedical fields (Sidey-Gibbons, 2019).

ML models can also be classified into grouping models (division of instances with respect to groups or classes) and gradient models for differentiation between instances or classes. Both generate an algorithm with respect to their own deductions. Some types of algorithm training are:

- Supervised learning. In supervised learning, a function is deduced from training data. The data for this type of training is made up of pairs of vectors (objects), where one part of the pair is the input data and the other is the output data (desired results). The output of the algorithmic function can be presented in numerical values (regression problems) or in a classification label. The objective of this type of learning is to develop a function with the ability to predict the appropriate value for the input object after visualization and analysis of a series of training data. For this type of learning to work properly and to be used as a reliable tool, it has to be exposed to a series of data presented to previously unseen situations. Experts in the field usually achieve this through training (Ang et al., 2015).An example of this type of algorithm is the one described in this thesis, where tumorclassification in WSI was intended. In broad terms, the learning model deals with classifying a series of vectors with respect to various categories (classes) or labelexamples. This type of learning has proven to be very useful in biological and medical research, being the basis of bioinformatics bioinformatics (Larranaga et al., 2006).

- Unsupervised learning. In this type of learning, the model is adjusted to the observations and the whole process is carried out with a set of examples formed only by inputs to the system, i.e., it is not fed with information about the classes. In addition, the system has as its principal job the automated recognition and classification of patterns to label new inputs. The training does not require experts for its development, so there is no pre-existing knowledge and it has the ability to self-organize. The network automatically discovers different features, regularities, correlations and categories in the input data. In a broad sense, unsupervised learning usually uses the input objects as a set of random variables to build a density model and a dataset. Another form of unsupervised learning is clustering, which can neglect probability methods. One of the advantages of this type of learning is that it requires less training time than supervised learning (Ang et al., 2015; Donalek, 2011).
- Semi-supervised learning. This is the mixture of the previous two and uses labeled and unlabeled data to classify them (Ang et al., 2015).
- Reinforcement learning. In this case, the algorithm learns with respect to the observation of the external environment, i.e., the information it uses to generate an algorithm is composed of the feedback it acquires from the outside world (trial-and-error principle). Since it does not require complete supervision, the autonomous training of the model only requires positive or negative reinforcement (punishment) that are derived from the good or bad performance of the model. The goal of this type of training is to enhance the algorithm's ability to understand the environment and make appropriate decisions to solve or understand problems (Sutton, 1992).

Not surprisingly, these types of learning resemble the way we humans learn. For us, this process is so automatic and simple that sometimes we fail to notice it; however, in ML, the learning method must be defined from the beginning and the rest will just be a reproduction of a repetitive sequence.

After learning, regardless of the model, the creation of the algorithms is generated with the following classification techniques:

- Decision trees. In this type of learning, the creation of a decision tree is necessary for the resolution of a problem. Its main objective is the generation of a prediction model based on logic for the correct representation and categorization of successive instances. Decision trees can be said to generate diagrams of sequential decisions with their probable outcomes, as is often used in economics where the option that avoids a loss or produces an extra profit has a value. The ability to create an option, therefore, has a value that can be bought or sold (Navada et al., 2011).
- Association rules. These algorithms are characterized by the creation of relevant relationships between different variables to determine the instances that occur within a data set. This type of algorithm is used to find relationships between variables within very large data sets. Among the best-known methods are the a priori algorithm, the Eclat algorithm and the Frequent Pattern algorithm (Chen et al., 2006).
- Genetic algorithms. The process of natural selection within evolutionary algorithms inspires this type of algorithm, i.e., it relies on genetic bases such as mutation and crossover to create new classification groups (Grefenstette, 1993).
- Artificial neural networks (ANN). The biological behavior of neuronal connections in animals inspired ANNs, i.e., they designed to solve problems in a similar way than the complex brain would do. It is made up of an extensive network of links with different numerical weights that work together to develop an output stimulus (Krogh, 2008). These connections contain thousands to millions of neuronal units and have the ability to adapt with respect to their individual experience and share information with each other. Each artificial neuron is interconnected with many others through links in order to receive and analyze information to generate an output. With previous weight multiplication of its value, this output will then be shared to the next neuron, to create a reaction of inhibition or activation, depending on the data. The output can also modify or limit the result that will be transmitted to the next link or neuron (Zou et al., 2008). ANNs are the basis of deep learning(DL) (Schmidhuber, 2015) and have been shown to perform successfully in accomplishing a large number of tasks such as computer vision (Zhou & Chellappa,

2012) and speech recognition (Lim et al., 2000), which are difficultto solve using previous algorithms.

- Support vector machines (SVM). These algorithms are defined as a series of methods related to supervised learning for classification and regression. In this type of algorithms, training is based on a first training phase, where multiple examples are fed in the form of pairs with their respective solutions, as well as a second phase of use for problem solving. Here, a "black box" is created to eject an answer to a certain type of problem and to predict the categorization of new examples (Meyer & Wien, 2015).
- Clustering algorithms. This kind of algorithms include an unsupervised learning method that has been commonly used in statistical analysis. The analysis is based on grouping observations (vectors) into subgroups (clusters) so that the observations in each group resemble each other according to criteria established by the algorithm's creator. Practically, they look for similarities within groups and separation of those that do not have similar characteristics (Fung, 2001).
- Bayesian networks. This one allows the creation of probability models represented in a series of random variables and their independencies through a directed acyclic graph. It combines observed evidence with "common sense" to determine the probability of presentation of occurrences with elements that are not necessarily linked to each other. An example is the creation of algorithms for determining the relationship of general symptoms to specific diseases. The results are generated graphically with the probability and conditions under which an instance would occur (Kotsiantis et al., 2007).

1.2.3 Deep learning

Finally, the subspecialty of AI within ML with which the most efficient software solutions have been developed in the medical field is deep learning (DL) (Suzuki, 2017). This is defined as a collection of ML algorithms that shape abstractions of large amounts of data using computational architectures that perform multiple nonlinear and iterative transformations of data expressed in a matrix form (Lee et al., 2017). DL is part of a broader group of ML methods based on resembling data representations. For example, an image that can be called an "observation" is represented in different forms, usually in a vector of pixels (depending on the type of data it is fed with) that enter into a network of analysis and classifications with respect to previously assigned examples or labels.

DL includes three types of architectures for algorithm creation (Shrestha & Mahmood, 2019):

- Deep neural networks (DNN),
- Deep convolutional neural networks (CNN) and
- Deep belief networks (DBN).

Although there is no single definition of a concrete DL algorithms, there is one point that all types of networks share. This point is centralized in the use of a cascade of layers with nonlinear processing units with the objective of extracting and transforming multiple variables. Each layer uses the output of the previous layer as input (in both, supervised learning or unsupervised learning) to finally model data and recognize patterns. Another feature shared by all types of artificial networks is learning based on multiple levels of features or data representations. Higher-level features are derived from lower-level features to form a hierarchical representation. This involves different levels of abstraction to generate a hierarchy of concepts and features at each layer.

There is no clear definition about the number of layers (transformations) that makes an algorithm to be considered as deep, but most researchers in the field consider DL to involve more than two intermediate transformations (Shrestha & Mahmood, 2019), which distinguishes it from the shallow learning.

1.2.4 Image recognition

Recognition is responsible for identifying and classifying objects in an image. Possibly one of its most common applications today is automatic image labeling, used for web content management and organization, but it is also useful for pattern recognition in a medical image. Thanks to a class of models known as CNN, image recognition has experienced formidable advances. Biological processes that take place in the visual cortex, where neurons recognize stimuli in a restricted area of the visual field and classify them with respect to stored information, inspire these models. This area partially overlaps with that of nearby neurons, collectively covering the entire visual field. As a result, CNN learn to respond to different image features (edges, shapes, etc.), such as the filter banks used in traditional and manually defined algorithms. In fact, the ability to learn such filters is one of the characteristic advantages of CNN, which in turn eliminates the manual effort required in feature design (Liu et al., 2017).

The learning algorithm of this type of network allows the extraction of the characteristics of each class from a previously classified training data set. To do so, it modifies the weights of the neurons that form the network and their values are iteratively calculated using the backpropagation method of supervised learning (Wu & Chen, 2015).

This backpropagation algorithm consists of two main stages:

For each element of the training set, the class to which it belongs is calculated according to the values that the network weights have at that moment (Leonard & Kramer, 1990). In this way, the algorithm can determine how good the classification is through an error function, as well as by comparing the classification with respect to the class to which the object actually belongs. Once the error made has been obtained, the algorithm propagates backwards the neurons with weights that contribute enough to the classification of the input. With this iterative process, the weights are updated for optimization using gradient descent algorithms (Kishore & Kaur, 2012). Subsequently, this method updates the weights of the network in the opposite direction of the gradient of the error function.

For a correct training of the CNN, it is necessary to specify a set of training data to define them. This requires a large number of labeled images of the object categories to be classified within a complete image. From these images, the CNN manages to obtain and collect specific characteristics of each class to learn to differentiate them from each other; the greater the number of training images, the better the results in the classification of new objects (Xin & Wang, 2019).

In either case, a set of labeled images must be chosen in which we have as many classes as different objects we want to classify. An additional class is also necessary for objects that do not fit into any of the training data (background). In this category, we must have as much data as possible and it works better the more diverse they are. This will allow the algorithm to avoid misclassifications and false positives in the detection of an object. In an image, the network as 'background' should classify everything that is not an object.

1.2.5 Evaluating the algorithm: confusion matrix

The evaluation of the algorithm performance is performed through a confusion matrix (CM). This method allows the visualization and analysis of the accuracy during the classification of objects even with a large amount of data. This type of matrices is usually used in supervised learning types.

Ting defined CM as a table that contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix (Ting, 2010).

Each column of the matrix represents the number of predictions in each class, while each row represents the instances in the actual class. One of the benefits of confusion matrices is that they make it easy to see if the system is confusing two classes (table 1).

The following table shows the CM for a two-class classifier taken from (Kulkarni, Chong, & Batarseh, 2020):

Prediction			
Ground truth (original label)		Negative	Positive
	Negative	True negative (TN)	False positive (FP)
	Positive	False negative (FN)	True positive (TP)

CM: Confusion matrix.

The entries in the CM have the following meaning:

- TN is the number of correct predictions that an instance is negative,

- FP is the number of incorrect predictions that an instance is positive,
- FN is the number of incorrect of predictions that an instance negative, and
- TP is the number of correct predictions that an instance is positive.

With these terms, accuracy, recall and precision can be calculated. The terms will be pointed in the next paragraph (Kulkarni et al., 2020; Provost & Kohavi, 1998; Ting, 2010).

- The precision also known as positive predictive value is defined as the proportion of the total number of predictions instances that were correct classified.
- The recall, also known as true positive rate or sensitivity is the quantity of positive cases that were correctly classified.
- The FP rate is the amount of negatives instances that were incorrectly identified as positive.
- The TN rate also known as specificity is defined as the quantity of negatives cases that were appropriately classified.
- The FN rate is the amount of positives instances that were incorrectly identified as negative.
- F1 Score is the measure of accuracy that has an artificial recognition model and is used in the determination of a single weighted value of accuracy and recall.

1.3 Computer-aided diagnosis

1.3.1 Overview and definitions

Computer-aided diagnosis (CAD) are medical procedures that support medical doctors in the interpretation of multimedia content obtained from tests that the patient has been subjected to, e.g., medical images (Giger & Suzuki, 2008). The idea of CAD is not to give a complete diagnosis from the original source, but to help the clinician who is writing the diagnosis to achieve an optimum diagnosis.

With this technology, the clinician is able to interpret all the visible information, since the machines process the whole picture and do not ignore any minor information that would otherwise escape the human eye. In this way, by highlighting the relevant information, they help the specialist not to overlook any detail.

CAD systems are an interdisciplinary technology, which is still very new at present, combining artificial intelligence, digital image processing and other subfields of medicine such as radiology, tomography or pathology. Image processing with the aid of complex pattern recognition systems makes it possible for the medical doctor, usually a radiologist, to interpret the information contained in the medical image with much lower difficulty (Doi, 2007).

CAD systems employ algorithms to analyze and recognize patterns in patient data that suggest possible anomalies. In a similar manner that a clinician is trained to identify anomalies by studying cases, CAD algorithms are taught to recognize patterns from an initial finite database with and without anomalies. This database is known as the "training set" (Chan et al., 2020).

Once the CAD system has been trained, it is then ready to be used on new patients to detect matching or discarding patterns of diseases or lesions. The pattern classifications in CAD devices are intended to be sufficiently reliable and efficient to support the specialist in identifying and diagnosing them. In general, device reliability is estimated using a different database known as a "test set" (Chan et al., 2020).

The methodology of CAD systems is very similar to that of a standard pattern recognition system:

- Preprocessing. In this step, all image imperfections such as noise are corrected and the image is harmonized in case of differences in exposure levels at different points.
- Segmentation. With the help of a database, matches are searched to detect important structures in the image and define them as regions to be analyzed individually.
- Structuring. Each of the previously defined regions is analyzed to extract important information from each of them regarding, for example shape, size, location, and so on.

At the end, defined regions that could be interesting for our diagnosis are left. The different regions that were previously identified as relevant are analyzed by means of several techniques. Each of these procedures has a limit that the region in matter must surpass to be considered relevant, if it is, the same procedure highlights it so that it does not go ignored by the specialist. It is the specialist who will finally decide what is relevant for the diagnosis and what is not, removing the latter for future consultations.

CAD systems are currently unable to detect 100% of pathological alterations. Thereliability (guess/hit rate) of these systems can reach up to 90% depending on the systemand the application (Xing et al., 2021). An incorrect guess, understanding as incorrect guess all those points that the system has marked as important without necessarily being so, are called false positives (FPs), so that the fewer FP we have, the more specific our procedure will be.

As with all technologies, CAD systems also have their own limitations (Fujita, 2020):

- Guarantee. There is no guarantee of a solution, it means, contrary to general thinking, that if the procedure works perfectly, it should not guarantee a diagnosis. All it guarantees is an image with areas for the relevant algorithm that the specialist will then have to consider.
- Annotations. Requires fine grained expert annotations.
- Database maintenance. The biggest problem with this technology is database maintenance. The algorithms that must detect the regions of interest need to consult some databases where the relationships between different cases are entered, a fact that causes the computational cost of the process to grow exponentially to prohibitive limits.
- High cost. As in most new technologies, the economic cost of CAD systems is still very high nowadays.

1.3.2 Computational pathology

The integration of AI and its subfields (e.g. ML) in medicine has accelerated the growth of different areas of medicine, mainly in imaging (Doi, 2007) and is now one of the central research subjects in digital pathology and toxicology (Turner et al., 2020). Conversely, due to the scarcity of accessible WSI databases, functional and standardized software, burden of annotations and validation of algorithms, the point of full and efficient development of automated diagnostics in oncology has not been reached yet (Chenet al., 2021).

The integration of DP and AI has opened the door to a completely new world ofpossibilities in diagnostic histology and research, into what we know as computer-aided pathology (CAP), or more specifically as computational pathology (CPATH). The expertsof the Digital Pathology Association (DPA) define CAPTH as a branch of pathology that involves computational diagnostic systems or set of methodologies that use computer programs to interpret pathological images (WSI), extract patterns and analyze patient specimens for the study of disease (Abels et al., 2019; Nam et al., 2020).

1.3.3 Previous work in human pathology

AI is already positioned as a fundamental tool for optimizing and automating mechanical tasks that require the prior analysis of a large amount of data in various sectors. Specifically, in the field of pathology it is already being used successfully to develop new therapeutic alternatives, accelerate molecular diagnostics or collaborate in clinical decision-making. Furthermore, in different diagnostic centers and research institutions, AI is already routinely helping medical professionals to diagnose diseases, plan personalized treatments or even design drugs for specific applications (Litjens et al., 2016).

The successes achieved so far point to a not-too-distant future in which medical doctors, AI and robots will work together in a coordinated way every day to apply a more precise and efficient medicine, thanks to overcoming human limitations in processing huge amounts of data. In that sense, AI does not need to be perfect to be useful in medical practice, it just needs to be better and faster than medical doctors in providing a diagnosis.

AI could help as well, to reduce the misdiagnosis rate in several types of diseases for example:

- Automated analysis and detection of prostate cancer in H&E slides (Bulten et al., 2018; Tolkach et al., 2020).

- Identification of breast cancer metastases in sentinel lymph nodes (Steiner et al., 2018; Wang et al., 2016).
- Automated grading of gliomas and astrocytomas (Ertosun & Rubin, 2015; Kolles et al., 1995).
- Ki67 Scoring (Narayanan et al., 2018).
- Genetic Mutation Prediction (Schaumberg et al., 2017).
- Differentiation between benign and malignant tumors, e.g., carcinoma vs. non-carcinoma (Bejnordi et al., 2017; Babak Ehteshami Bejnordi et al., 2018; Yahui Jiang et al., 2020).
- Colon cancer classification (Awan et al., 2017; Kainz et al., 2017).
- Gastric cancer classification (Shujun Wang et al., 2019).
- Tumor subtyping (Yun Jiang et al., 2019).
- Mitotic count (Veta et al., 2015).
- Evaluation of biomarkers (Khameneh et al., 2019; Vandenbergheet al., 2017).
- Lung cancer classification/subtyping (Coudray et al., 2018; Gertych et al., 2019; Teramoto et al., 2017).
- Count of immunologic cells (Aprupe et al., 2019).
- Prognosis prediction (Shidan Wang et al., 2018).
- Classification of melanocytic lesions (Norgan et al., 2018; Wang et al., 2020).
- Classification for bone marrow aspirate differential counts (Chandradevan et al., 2020).

...and the list goes on. However, there are not that many studies using AI, specifically DL in veterinary pathology.

Some of the most relevant studies are mentioned below.

1.3.4 Previous work in veterinary pathology

As mentioned above, there are not many publications that focus on the development of algorithms for routine diagnosis in veterinary medicine. However, the development of some complementary tools for the diagnosis of animal diseases, mainly in surgical and toxicological diagnosis, has been described (Zuraw & Aeffner, 2022).

Some highly relevant advances are mentioned below.

- Mitosis detection in dogs (Bertram et al., 2021; Bertram et al., 2020).
- Pigment quantification within cells in horses (cytologic slides) (Marzahl et al., 2020).
- Differentiation of round cell tumors in dogs (Salvi et al., 2021).
- Classification of mammary tumors in dogs (A. Kumar et al., 2020).
- Retinal evaluation in mice (De Vera Mudry et al., 2021).
- Lung fibrosis and inflammation characterization in mice (Heinemann et al., 2018).
- Automatic glomerular identification and quantification in mice (Sheehan & Korstanje, 2018).

1.4 Algorithm development in pathology: establishment of ground truth

Regardless of the type of hardware and software chosen by the user, the implementation of algorithms based on supervised ML has great advantages in the clinical and histopathological diagnostic workflow, as once they are properly developed and validated by computer engineering and pathology experts, they can improve the accuracy, speed and efficiency of diagnosis. This allows pathologists and researchers to analyze features on slides that are not easily identified with conventional optical evaluation; in addition, it also allows them to make a more complete evaluation of the slide in a short time, as well as detect and classify structures of interest and gather objective data (Abels et al., 2019; Aeffner et al., 2017). However, to enjoy these benefits, accurate and extensive convolutional neural network (CNN) training is required.

The first step to achieve the development of a reliable algorithm in supervised DL (besides the raw image sets) is the establishment of a ground truth. Since this process is the basis of all DL training, it is also the most time-consuming and challenging (Irshad et al., 2015). The establishment of a ground truth is based on the labeling of specific features within a slide. This can be done on previously selected patches or on the entire digitized slide (Dimitriou et al., 2019). Naturally, training neural networks on a completeslide with high resolution requires a great amount of time, effort and expertise. Since it isa supervised training, the medical expert (pathologist in this case) carries out the labelingof each structure manually, practically it is a matter of transforming the visual experienceand skill into computational data with coordinates and pixels that will later enter the CNN. Depending on the type of algorithm to be developed, the labeling of structures can be carried out with different tools (depending on the software used); for example, specific zones contained in geometric figures (circles, squares, rectangles), zones surrounded by lines (polygons), or even automatic delimitations that follow the lines between tissue or cellular structures (magic wand) (Aubreville et al., 2018).

Obtaining adequate datasets for DL can become a difficult task for the expert pathologist due to the tedious nature of the task (Dimitriou et al., 2019; Irshad et al., 2015). Once the ground truth is defined and correctly annotated by the expert pathologist, the data can enter to the convolutional neural network to train the final algorithm. Currently, it is difficult to establish quality control at this step; however, the most successful way to achieve this is to divide the database (slides) into 3 sets: training set, validation set and test set (Abels et al., 2019). Once the method is validated, the algorithm is applied to the test set and compared with the ground truth, or pathologists' diagnosis. The definition of a gold standard for the ground truth is controversial, as the results often fall into the "gold standard" paradox (Aeffner et al., 2017), where the algorithm data ends up being more reproducible than the human ones. One of the ideal methods to avoid this and evaluate the algorithm more accurately would be to compare the algorithm results again with new expert opinions and even use other immunomolecular tests such as IHC; however, this is not often accomplished. Finally, after evaluation of the accuracy of the algorithm and its validation, the reproduction and implementation are ready to be passed on.

2. Basic consideration and working hypothesis

The use of ML in DP has proven to have useful applications, such as in mammary carcinoma diagnosis and metastasis detection (Araujo et al., 2017; B. Ehteshami Bejnordi et al., 2017; Sudharshan et al., 2019), automated mitosis detection (Bertram et al., 2021; Bertram et al., 2020; Roux et al., 2013), melanocytic skin tumor classification (Norgan et al., 2018), quantitative evaluation of immunostaining (J. X. Liu et al., 2019), round cell tumor differentiation (Salvi et al., 2021) and so on. However, to date no algorithm has been developed that determines or discriminates tumors of completely different tissue origins in H&E-stained WSI.

Because the skin is one of the most common anatomical sites of neoplasia in dogs and canine skin tumors represent the largest number of cases in veterinary pathology diagnostic centers (Dorn, 1967; Dorn et al., 1968; Gamlem et al., 2008; Merlo et al., 2008), we decided to conduct a study focused on creating an algorithm that would be able to automatically classify and identifyseven of the most important and common canine tumors (Graf et al., 2018; Kok et al., 2019).

We hypothesis that the training of an artificial neuronal network using an appropriate number of well annotated digital images of canine cutaneous tumors will lead to a software solution, which identifies and differentiates common canine cutaneous tumor types with a similar sensitivity and specificity as a trained pathologist.

3. Material and methods

3.1 Case selection and scanning

Surgical biopsies of seven of the most frequent tumors in dogs were retrospectively selected from the histopathology archive of the Institute of Veterinary Pathology of the Free University of Berlin. The tumor types were trichoblastoma, squamous cell carcinoma (SCC), melanoma, peripheral nerve sheath tumor (PNST), mast cell tumor (MCT), plasmacytoma and histiocytoma. First, 50 cases per tumor were chosen with respect to typical histological features, state of preservation, sufficient histological perceptibility of cellular details and staining quality (H&E, total n=350 cases/slides). All glass slides were digitalized to generate WSIs using a linear scanner (ScanScope CS2, Leica) in 1 focal plane by default settings and they were scanned at a magnification of 400× (image resolution: 0.25 µm/pixel). These WSI entered to our first dataset.

Additionally, in a similar fashion, we chose another 20 slides per tumor type, which were previously diagnosed and reviewed with typical features of each tumor (n=140 cases/slides) to scan with the same method. These WSI entered to the second dataset.

3.2 Dataset

The first 350 slides were included in Dataset 1, with the aim of including them in training, validation and testing (method described below) (Wilm et al., 2022). The following 140 slides (dataset 2) were included only as a test set for the algorithm and for the "human vs machine challenge".

3.3 Annotations of tissue area and tumors

Annotations were performed in SlideRunner (Aubreville et al., 2018), a software for massive annotations in WSI. This software was developed by working group at the FU Berlin together with IT-specialists of the Friedrich-Alexander-Universitaet (FAU) Erlangen. We completed annotations in the WSIs of the first database (350 WSI/50 per tumor type). Annotations were made using the polygon tool, surrounding each area of interest with a thin line, from point to point until it is completely delimited as precisely as possible (Figure 1). The corresponding class was assigned to each surrounded area by the points of the polygon tool. The annotation classes were: epidermis, dermis, subcutis, trichoblastoma, melanoma, PNST, SCC, MCT, histiocytoma, plasmacytoma, inflammation/necrosis, bone, cartilage. All the tissue structures of the slide were annotated with the aim of trainingthe algorithm as accurate as possible.

A total of thirteen classes were created, focusing on the main histological structures of the skin (epidermis, dermis, subcutis), as well as the seven tumor types. In addition, special annotations were made for miscellaneous tissues that were not necessarily relevant to this study but were found in some cases such as inflammation, necrosis, bone and cartilage. Table 1 shows the total number of annotations per class, as well as the total surface area in mm^2 annotated per class on SlideRunner.

No annotations were performed in WSI of the second dataset.

Table 1. Total number of annotations created in SlideRunner and the total annotation area for each of the classes in mm^2. SCC: squamous cell carcinoma; PNST: peripheral nerve sheath tumor.

Annotated class	**Annotations**	**Annotation area (mm^2)**
Epidermis	3188	2244.57
Dermis	3423	16616.21
Subcutis	2850	7369.88
Trichoblastoma	423	9072.1
SCC	337	3542.28
Melanoma	379	6836.93
Plasmacytoma	377	4750.34
Mast Cell Tumor	161	9330.1
PNST	131	11108.78
Histiocytoma	369	2947.59
Bone	51	216.86
Cartilage	16	32.15
Inflammation/Necrosis	719	2050.16
Total of annotations	**12424**	**76118.05**

MCT: mast cell tumor; PNST: peripheral nerve sheath tumor; SCC: squamous cell carcinoma.

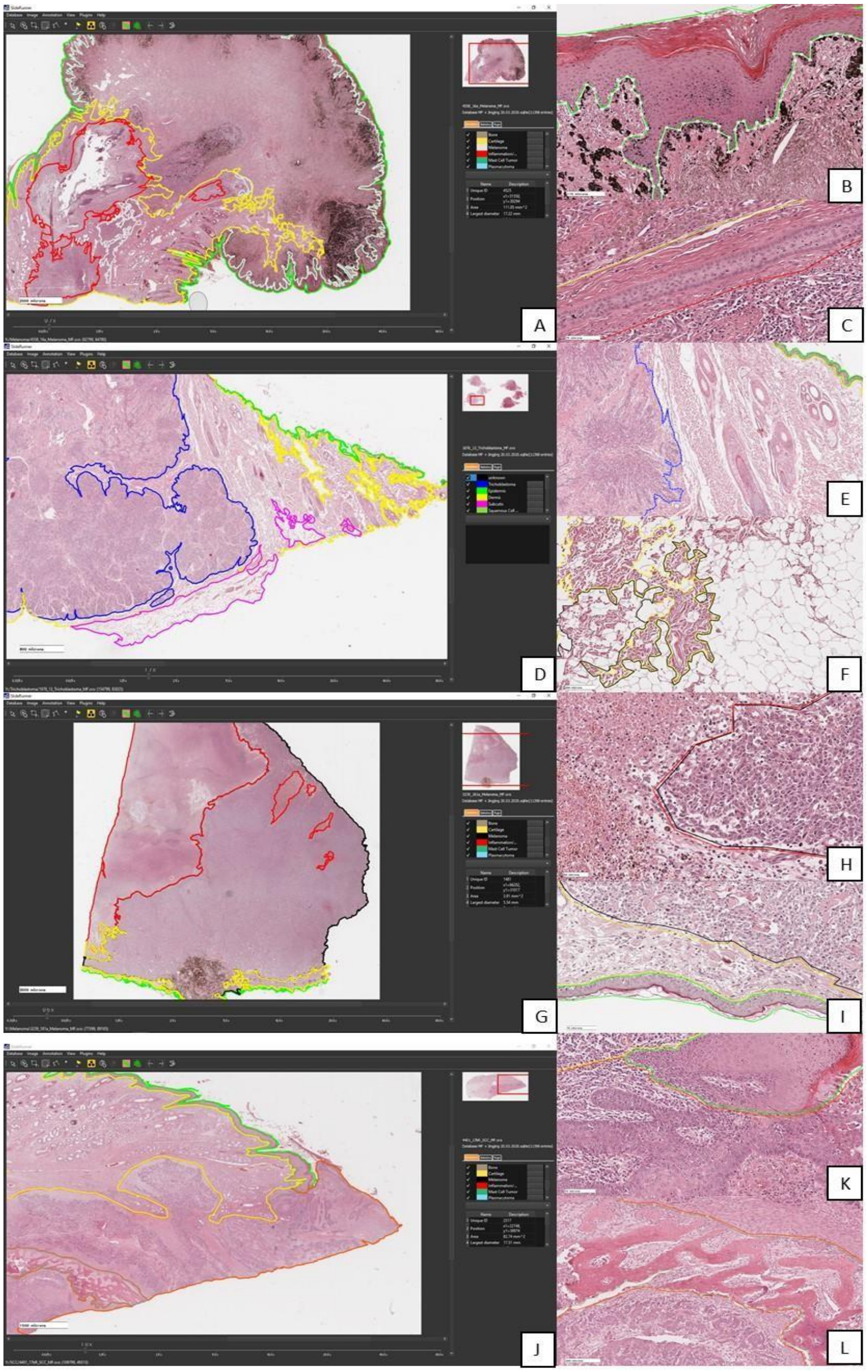


Figure 1. Annotations performed in WSI in SlideRunner with the polygon tool.

A. Melanoma, skin. Annotations completed in WSI with the polygon tool. Colored lines surround the desired structures.

B. Magnification of the superficial region of the tumor, the division of the epidermis and the tumor is observed.

C. Magnification of the area between the tumor and subacute inflammation. In the middle of these two is the dermis with a normal follicle and collagen. Green: epidermis; yellow: dermis; red: inflammation/necrosis; white: melanoma. WSI, HE.

D. Trichoblastoma, skin. Annotations completed in WSI. The edges of the tumor are delineated by a blue line.

E. Magnification of the region showing annotations made with clear demarcation between the tumor and the dermis.

F. Magnification of the subcutaneous region during annotation of the subcutaneous tissue. Note the collagen of the dermis surrounded by a yellow line and the black line shows the unfinished annotation process.

G. Melanoma, skin. Completed annotations in WSI show the tumor surrounded by a black line and a central area of necrosis surrounded by a red line. Four additional random regions of necrosis within the tumor were also annotated.

H. Magnification of an annotated necrosis zone next to a tumor region. Note the poor demarcation between both tissue types.

I. Magnification of the superficial zone of the tumor where the difficulty in delineating the dermis of the tumor and the epidermis can be observed. The annotations were made as precise as possible.

J. Squamous cell carcinoma (SCC), skin of the paw. Annotations completed in WSI. The tumor is shown delineated by an orange line and at the bottom-center of the tumor, a region composed of bone, delineated by an olive green line.

K. Magnification of the transition zone between the epidermis (green line), the SCC and the dermis (yellow line). Note the difficulty of this task in attempting to define borders between these three histologic structures.

L. Magnification of the region composed of bone surrounded by the olive green line and its poorly demarcated vicinity with the SCC (orange line).

3.4 Development of the algorithm, training and testing

Technical method development was conducted at the pattern recognition laboratory of the Friedrich-Alexander-University Erlangen-Nürnberg in close collaboration with the medical experts at the Freie Universität Berlin (Wilm et al., 2022).

From the first dataset, the 50 images of each tumor type were divided into 35 training, five validation, and 10 test images. For the total dataset, this resulted in 245 training, 35 validation, and 70 test WSIs. We trained a neural network for the segmentation into six classes: background, tumor, epidermis, dermis, subcutis, and inflammation combined with necrosis. For this, we chose a UNet (Ronneberger et al., 2015) architecture with a ResNet18 (He et al., 2016) backbone pre-trained on ImageNet (Russakovsky et al., 2015).

We trained the network with image patches sized 512 x 512 pixels and a resolution of 4.0 µm/pixel. Due to high class-imbalances, we followed an adaptive sampling strategy. Initially, ten patches per slide were sampled uniformly across all annotation classes, resulting in 2,450 training patches. These were used to train the network for one epoch. Then, the network performance was evaluated on 350 validation patches (10 per WSI) sampled in the same fashion. Afterwards, the probability of sampling patches from a class with a low validation performance was increased, whilst high-performing classes were under-sampled. By using this adaptive sampling scheme, we explicitly trained the model on difficult classes, aiming for faster convergence of the model training. We trained the model for 100 epochs with a maximal learning rate of 10^{-4} and a batch size of four. As loss function, we used a combination of cross-entropy and dice loss.

Additionally, we trained a tumor type classification network to distinguish between the seven tumors. We used the same dataset split as used for training the segmentation network, resulting in 35 training images per subtype. Due to the high morphological resemblance of round-cell tumors, which might only be distinguishable at a high image resolution, we decided to train the classification network on patches at the original resolution of 0.25 µm/pixel. To cover as much context as possible, we increased the patch size to 1024 x 1024 pixels. We used an EfficientNet-B5 (Tan & Le, 2019) architecture pre-trained on ImageNet (Russakovsky et al., 2015). For each epoch we sampled 10 patches per slide, ensuring a uniform sampling across all tumor types. We additionally trained the network on a "non-neoplastic" class which was trained on patches from all remaining annotation classes (epidermis, dermis, subcutis, and inflammation combined with necrosis). A patch was only used for training the classification network if at least 90 % of the pixels were annotated as the sampled class. We used a batch size of four and a maximal learning rate of 10^{-3} and trained the network for 100 epochs until convergence. For optimization, we used the cross-entropy loss and the Adam optimizer.

Figure 2 visualizes the WSI inference pipeline. A slide was first segmented into six classes using the segmentation network. Afterwards, regions segmented as tumor were classified into one of the seven tumor types. For this, we upscaled the predicted tumor region from the segmentation resolution of 4 μm/pixel to the classification resolution of 0.25 μm/pixel. Then, we divided the tumor region into patches sized 1024 x 1024 pixels, which were only passed on to the classification network if they were completely segmented as tumor. Each patch then obtained a classification label and all patches classified as non-neoplastic tissue were excluded. All remaining patch classifications were combined to a slide classification label using majority voting (Wilm et al., 2022).

In the same way, the algorithm was run on the second dataset, which did not contain annotations, only hidden diagnostic labels for each WSI.

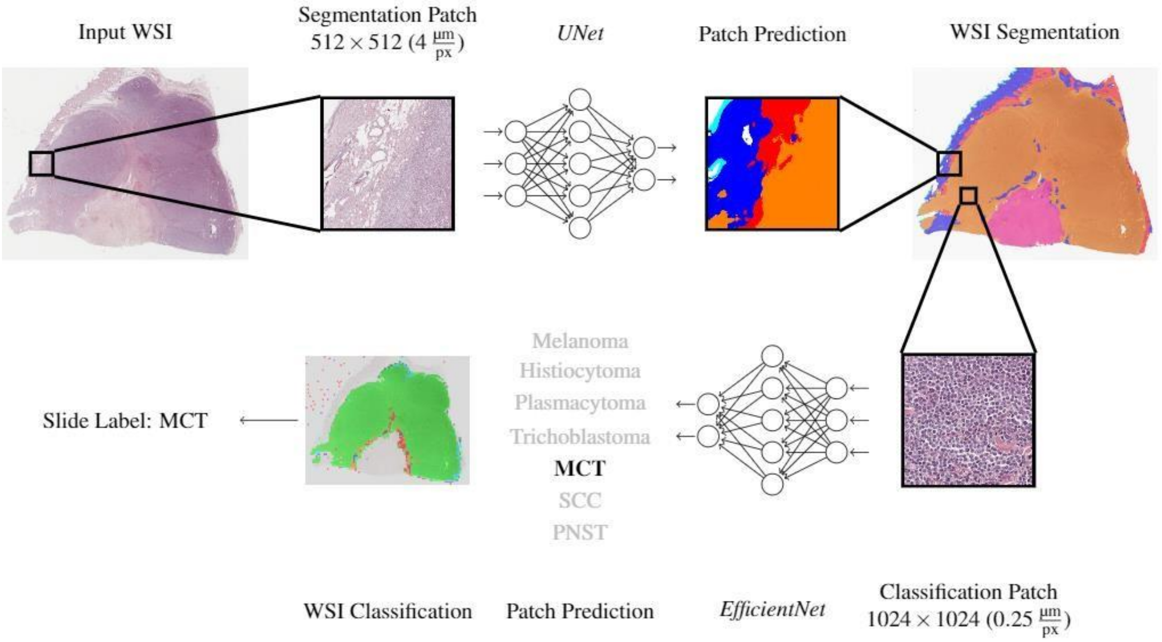


Figure 2: Cutaneous tumor segmentation and classification pipeline.

3.5 Human vs machine challenge

In order to compare the results of the algorithm with human intelligence, a challenge was carried out, which consisted of providing the second dataset containing 140 WSI (20 per tumor type) to 6 experienced board-certified pathologists for their evaluation. Since this is a comparative experiment, each participating pathologist was asked to assign each slide a main diagnosis and two differential diagnoses. The only condition was that it was not possible to assign a percentage less than 1% or greater than 98% to each diagnosis and that in total they should result in 100%. The percentage assigned by the pathologists to each response or diagnosis was defined as confidence (certainty/sureness).

In addition, in order to perform a uniform statistical analysis for this experiment, we took into account the final classification of the algorithm and from the total patch count, only those first three options. Finally, we transformed the count of only the first 3 options into 100% of the patches. The accuracy of the pathologists and the algorithm were similarly evaluated.

We first defined accuracy as the final diagnosis that was chosen by both the pathologists and the algorithm for each slide qualitatively, i.e. regarding the algorithm, the tumor type with the highest number of classified patches on a slide was considered as the main diagnosis and regarding the pathologists' answers, the tumor type for which the majority of pathologists voted was considered as main diagnosis.

Subsequently, precision (positive predictive value) was defined as the total number of true positive percentages assigned over the total of the remaining percentages of their differential diagnoses. In the case of the algorithm, it was defined as the total number of patches correctly classified over the total number of patches collected.

Recall (sensitivity) of the pathologist’s answers was defined as the total percentage of correctly diagnosed cases over the fraction of true positives and false negatives. In the case of the algorithm, it was defined as the total number of patches correctly classified over the total number of true positives and false negatives.

In addition, the average of the percentages assigned by the pathologists for these calculations and the average of pathologists who agreed on the diagnoses were collected.

The responses of the six pathologists for each slide were averaged and displayed as whole numbers or decimals (depending on the result). For example, if on one slide 4/6 pathologists chose one tumor type and 2/6 chose another type, the final number of accuracy would be 4/6; if on the next slide of the same tumor type 5 pathologists chose one tumor type and 1 pathologist chose another, the average correct answer would be 4.5 (sum of 4 + 5 / 2), and so on until completing the 20 slides per tumor type (140 WSI).

Statistical comparison of this section was conducted at the Institute of Veterinary Epidemiology and Biometry at the Freie Universität Berlin.

3.6 Immunohistochemistry

In order to reconfirm the most confounding cases in our database (second dataset) and determine their individual ground truth, a conventional IHC analysis was performed. The following antibodies were used: Melan-A (A103) monoclonal mouse for melanoma (de Wit et al., 2004; Ohsie et al., 2008; Ramos-Vara & Miller, 2011) in a 1:300 dilution. CD79acy monoclonal mouse anti-human for plasmacytoma (Baer et al., 1989; Ramos-Vara et al., 2007), in a 1:60 dilution. CK10 (EP1607IHCY) monoclonal rabbit for SCC (Assawawongkasem, et al., 2020) in a 1:1000 dilution. E-Cadherin (EP913 (2) y) monoclonal rabbit for histiocytoma (Baines et al., 2008) in a 1:1000 dilution.

4. Results

4.1 Dataset

From the first dataset, a total of 350 slides were scanned and fully annotated. In total, 12,424 annotations were made, with an annotation surface of 76,118.05 mm^2 (Table 1, Figure 1). Bone and cartilage annotations were excluded in the training due to their low diagnostic relevance and their low number of annotations and annotated area in our database. Once the model was trained and validated, we first tested it on the test split of dataset 1 and evaluated the segmentation output against our ground truth annotations. Subsequently, we applied the algorithm to the second dataset (140 WSIs) and compared the tumor-type classification performance to the six pathologists.

4.2 Algorithm performance: tissue segmentation

In the first model run in the first dataset, the performance of the algorithm in segmenting tumor vs. non-tumor classes was evaluated. Within the non-tumor classes, we only considered dermis, epidermis, subcutis and inflammation/necrosis. As mentioned above, we excluded bone and cartilage classes in our model because of their negligible annotation count and low diagnostic relevance in cutaneous oncologic pathology. The precision of the segmentation model was 78%; however, the precision regarding tumor segmentation and exclusion of those patches that did not contain tumor (i.e., tumor vs. non-tumor) was 95%. Excluding tumor segments, the class with the best accuracy was subcutis with 85%, followed by dermis with 84% and epidermis with 79%. The class with the lowest precision was inflammation/necrosis, with 46%. The confusion of this last class was mainly with the tumor class, which was incorrectly segmented in 26% of the patches. False positives were also confused with dermis in 15%, subcutis in 11% and epidermis in 2%. Similarly, the tumor class was the one with the highest number of false negatives, mainly confused with inflammation/necrosis (26%). Confusion between tumor and dermis and epidermis was 10% and 12%, respectively. The summary of the segmentation confusion matrix is shown in Table 2. The total number of patches that were considered as tumor (which had a precision of 95% and recall of 66%, F1-score 78%) entered the classification model. Figure 3 shows some performance examples of the algorithm for the segmentation of dataset 1 WSIs. The performance of the algorithm for tumor classification is described below.

Table 2. Segmentation confusion matrix.

Prediction						
Class		Dermis	Epidermis	Subcutis	I/N	Tumor
	Dermis	**0,844**	0,080	0,126	0,150	0,033
	Epidermis	0,009	**0,789**	0,001	0,022	0,001
	Subcutis	0,042	0,005	**0,854**	0,111	0,006
	I/N	0,004	0,008	0,004	**0,456**	0,008
	Tumor	0,102	0,117	0,015	0,261	**0,952**
Segmentation Precision		**0,844**	**0,789**	**0,854**	**0,456**	**0,952**
Segmentation Recall		**0,684**	**0,960**	**0,839**	**0,948**	**0,658**
F1 Score		**0,756**	**0,866**	**0,846**	**0,616**	**0,778**

I/N: Inflammation and necrosis.

4.3 Algorithm performance: tumor classification

In our second dataset, we used 140 WSI (20 per tumor type) to test the algorithm only regarding its performance by classifying tumor types. The slide-level accuracy (i.e., the final class/tumor type that was ranked highest with respect to the total number of patches on each WSI) was 95%, i.e. out of 140 slides, 133 slides were correctly classified. The patch-level precision (i.e., the correct ranking of the algorithm with respect to the individual patches and their summation over all WSIs) of the model was 85%. Table 3a shows an accuracy summary of our model at slide level (qualitative and definitive classification per WSI).

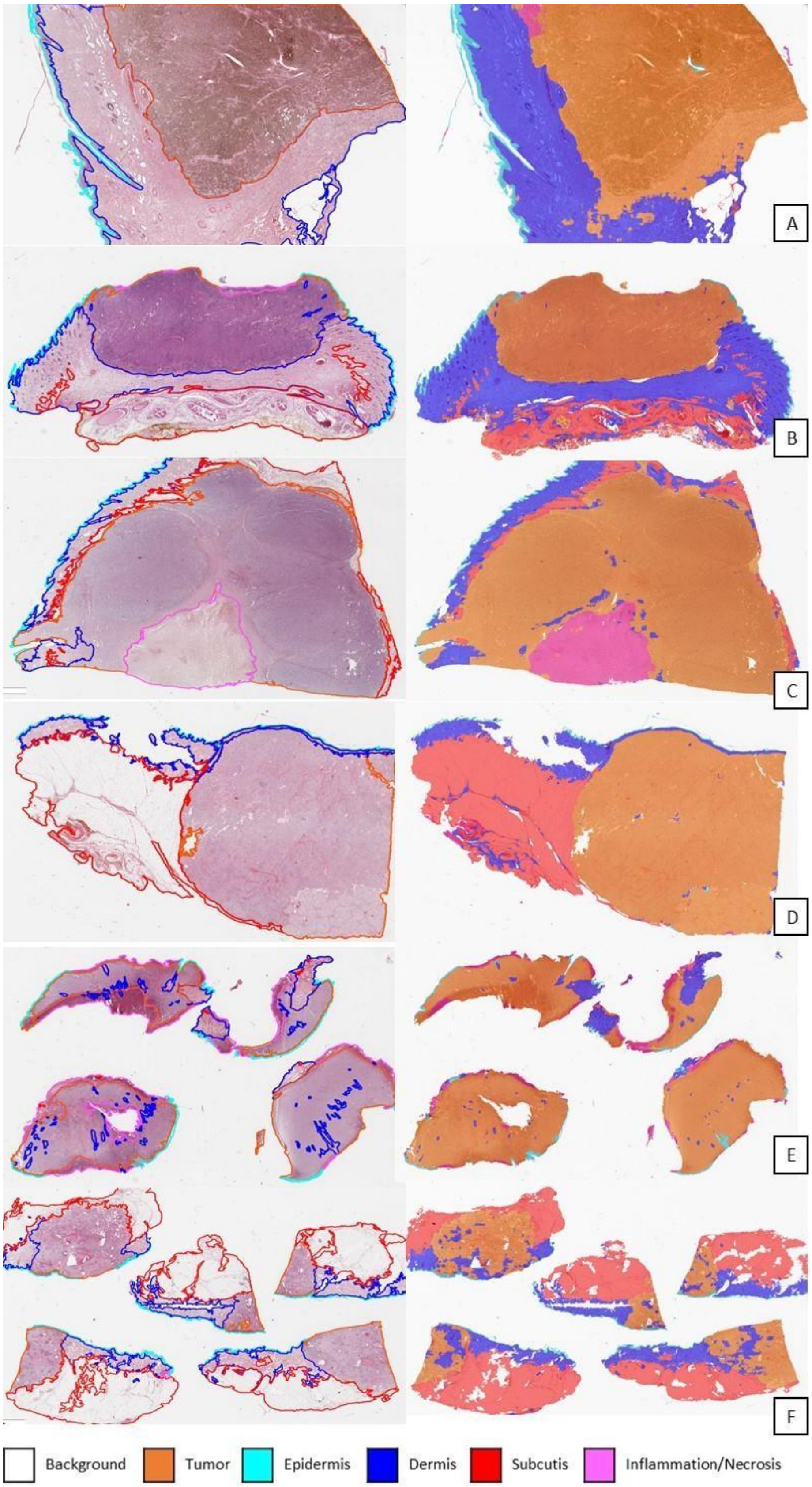

A
B
C
D
E
F
Background
Tumor
Epidermis
Dermis
Subcutis
Inflammation/Necrosis

Figure 3. Comparison of the annotations completed in SlideRunner (left side) and the performance of the segmentation algorithm on dataset 1 (right side).

A. Melanoma. Note the area of necrosis (pink) that was automatically segmented by the algorithm. Some parts of the dermis were segmented as part of the tumor (orange) due to its unclear division.

B. Histiocytoma. The algorithm ignored the annotated areas of ulceration. Also, note the efficiency in detecting fat within the dermis and its inclusion in the "subcutaneous" class (in red).

C. MCT with a central area of necrosis. Note the precision of the model in segmenting the zone of necrosis almost as the original annotation. Random areas segmented as dermis within the tumor are shown (blue).

D. Trichoblastoma. Excellent segmentation performance, with only some regions within the tumor segmented as dermis (blue).

E. Histiocytoma. Excellent segmentation performance, respecting the areas of ulceration and segmented within the "inflammation/necrosis" class (pink).

F. Squamous cell carcinoma (SCC). Note the difficulty in the demarcation of the tumor and its differentiation with the dermis. Regions of dermis within the tumor were segmented randomly; however, the tumor was delineated with very favorable performance.

Orange: tumor; light green: epidermis; blue: dermis; red: subcutaneous; pink: inflammation/necrosis. WSI, H&E, panoramic view.

Trichoblastoma and PNST were properly classified in all the slides (20/20; slide accuracy 100%), with a patch-level precision of 94% and 91%, respectively. Melanoma wasproperly classified in 19 slides (95% accuracy) with a precision of 91% and one was misclassified as PNST. MCT was properly classified in 19 slides (95% accuracy) with a precision of 95% and one slide was misclassified as SCC. Histiocytoma was properly classified in 19 slides (95% accuracy) with 80% of precision and misclassified as plasmacytoma in one slide. SCC was properly classified in 18 slides (90% accuracy) witha precision of 70% and misclassified as plasmacytoma in 2 slides. Plasmacytoma was properly classified in 18 slides (90% accuracy), with 75% of precision and misclassified inone slide as melanoma and in one slide as SCC. Table 4 shows the confusion matrix of our model.

Some illustrations of the automatic classification of our model, with their respective normal histologic images, can be seen in Figures 4-8.

Table 3. Accuracy of the algorithm at the slide level of the 20 WSI/tumor type of our model (a) and the consensus of the 6 pathologists (b). MCT: mast cell tumor; PNST: peripheral nerve sheath tumor; SCC: squamous cell carcinoma.

A

Prediction

Label

Tumor	Melanoma	Plasmacytoma	MCT	PNST	SCC	Trichoblastoma	Histiocytoma
Melanoma	19	0	0	1	0	0	0
Plasmacytoma	1	18	0	0	1	0	0
MCT	0	0	19	0	1	0	0
PNST	0	0	0	20	0	0	0
SCC	0	2	0	0	18	0	0
Trichoblastoma	0	0	0	0	0	20	0
Histiocytoma	0	1	0	0	0	0	19

Slide-Level Accuracy: 0.95

B

Pathologists consensus

Label

Tumor	Melanoma	Plasmacytoma	MCT	PNST	SCC	Trichoblastoma	Histiocytoma
Melanoma	18	0	0	1	1	0	0
Plasmacytoma	0	19	1	0	0	0	0
MCT	0	0	20	0	0	0	0
PNST	0	0	0	20	0	0	0
SCC	0	0	0	0	20	0	0
Trichoblastoma	0	0	0	0	0	20	0
Histiocytoma	0	0	0	0	0	0	20

Slide-Level Accuracy: 0.98

Table 4. Confusion matrix of our model run on the second dataset of 140 WSI at a patch level. MCT: mast cell tumor; PNST: peripheral nerve sheath tumor; SCC: squamous cell carcinoma.

Prediction

Label

Tumor	Melanoma	Plasmacytoma	MCT	PNST	SCC	Trichoblastoma	Histiocytoma
Melanoma	0,913	0,019	0,003	0,020	0,037	0,017	0,008
Plasmacytoma	0,049	0,748	0,013	0,011	0,058	0,013	0,053
MCT	0,016	0,032	0,954	0,019	0,087	0,002	0,085
PNST	0,003	0,039	0,010	0,908	0,071	0,006	0,017
SCC	0,015	0,025	0,009	0,032	0,703	0,020	0,027
Trichoblastoma	0,003	0,023	0,003	0,005	0,014	0,936	0,010
Histiocytoma	0,002	0,115	0,007	0,005	0,031	0,005	0,800

Tumor Precision	**0,913**	**0,748**	**0,954**	**0,908**	**0,703**	**0,936**	**0,800**
Tumor Recall	**0,898**	**0,791**	**0,799**	**0,861**	**0,846**	**0,942**	**0,829**
F1 Score	**0,906**	**0,769**	**0,870**	**0,884**	**0,768**	**0,939**	**0,814**

Patch-Level Precision 0.85

4.4 Human vs machine challenge

The same slides used in the test set (n=140 slides, 20 per tumor type) were used in human vs machine challenge. Qualitatively, the slide level accuracy of the pathologists was 98% (137/140), with an average of 5.8/6 correct answers (i.e. average agreement of six pathologist on each of the 140 WSIs, table 3b). These results were obtained qualitatively, counting only the majority of votes received by the different diagnoses (similar to the final decision of the algorithm) and ignoring the rest, regardless of their vote count (1-0). However, in order to understand the behavior and compare it with the predictions of the algorithm, the accuracy, precision and recall calculations were carried out, as well as the F1 score with respect to the percentages given by the pathologists (quantitatively).

Summary results of the precision (positive predictive value) and recall (sensitivity) of the pathologists' answers and of our model's classification are shown in Table 5. The comparison of both classifications (diagnostic) with the balanced values (F1 score) is shown in Table 6.

Statistical analysis of this section was conducted at the Institute of Veterinary Epidemiology and Biometry at the Freie Universität Berlin.

Trichoblastoma, PNST, SCC, MCT and histiocytoma were properly diagnosed in all slides with an accuracy of 99%, 100% 100%, 98% and 90%, respectively. Melanoma was misdiagnosed in 2 cases, in the slide number 98 as PNST with an average of 4/6 votes in favor of this tumor on all slides of this group and in the slide number 119 as SCC again with an average of 4/6 votes in favor of this tumor on all slides of this group. IHC test of the slides number 98 against Melan-A was strongly positive (Figure 6c); in the 119 was slightly positive, but the IHC against CK10 was negative. Plasmacytoma was misdiagnosed as MCT in the slide 35 with 3/6 votes for this tumor type of this group. This case had a strong positivity in IHC for antibodies against CD79.

4.4 Results by tumor type

The results of the algorithm and the pathologists grouped by tumor type are described below:

4.4.1 Trichoblastoma

All slides were successfully classified as trichoblastoma (100% accuracy), with a precision and recall of 94% each. Algorithm confusion occurred mostly with SCC and melanoma, with 2% and 1.7% respectively (false positives recovered). The model ignored false negatives that mainly classified as plasmacytoma in 2.3% of the patches and again with SCC in 1.4% (false negatives).

The six pathologists (6/6) primarily diagnosed all slides as trichoblastoma with a confidence of 97% (considering that this number was the maximum they could reach, as if it was 100% confidence). This shows the most homogenous results. As the second (differential) diagnosis, SCC was the most voted in all slides, with an average of 4/6 votes, with a confidence of 1%. Melanoma was the third diagnosis assigned by the majority of pathologists in the slides, with an average of 3.5/6 votes and a confidence of 1%. This means that from the options provided to the 6 pathologists, the majority considered the most likely differential diagnosis of trichoblastoma to be SCC and Melanoma as third option.

4.4.2 PNST

All slides were correctly classified as PNST (100% accuracy), with a precision of 91% and a recall of 86% (F1=88%). Confusion of predicted patches occurred in 3.2% with SCC and 2.0% with melanoma (false positives). Patches ignored by the algorithm that were classified as (false) negative, were classified as SCC (7%) and plasmacytoma (4%).

The six pathologists primarily diagnosed all slides as PNST, with a confidence of 96%. This shows the second most homogenous results. As the second diagnosis (differential), melanoma was the most frequently mentioned by pathologists (5.3/6 votes in average), with a confidence of 2%. As the third diagnosis, SCC was the most voted diagnosis in the majority of slides (16/20), with an average of votes of 3.2/6. Trichoblastoma was the most voted in the remaining slides (4/20), with an average of 2.5/6 votes and a confidence of 1%. This means that from the options provided to pathologists, the most likely differential diagnosis of PNST is melanoma and SCC as a third differential diagnosis.

4.4.3 Melanoma

The model appropriately classified 95% of the slides (19/20) and one of them was misclassified as PNST. The precision of this group was 91% with a recall of 90% (F1=91%), as it falsely recognized 5% of the patches as plasmacytoma, 2% as MCT and 1.5% as SCC. Likewise, the remaining patches were ignored as melanoma and identified as SCC in 4% and as PNST, plasmacytoma and trichoblastoma in 2% (false negatives).

The six pathologists diagnosed 18/20 slides as melanoma, with a confidence of 96%. In one slide, 4/6 votes of the pathologists was PNST as the primary diagnosis, with a confidence of 97%. The IHC of this case was positive for antibodies against Melan-A (A-103, figure 7C), so the confirmatory diagnosis is melanoma. Furthermore, from this slide 4/6 votes were for melanoma as the second diagnosis. In one slide, 4/6 votes were for SCC as the primary diagnosis, with a confidence of 91%. The IHC of this case was slightly positive for Melan-A and negative for CK10, so the confirmatory diagnosis is Melanoma. From this slide, the most voted second diagnosis was plasmacytoma. From this group, the most voted secondary (differential) diagnosis was PNST (4.4/6 votes) with a confidence average of 2% and the most voted third diagnosis was trichoblastoma (3.6/6 votes), with a confidence of 1%. This means that from the options provided, 4.4/6 votes of the pathologists agreed that the most likely differential diagnosis of melanoma is PNST, and 3.6/6 votes of the pathologists agreed that the third is trichoblastoma.

4.4.4 MCT

Similar to the previous one, 95% of the slides were correctly classified as MCT (19/20) and the remaining slide was incorrectly classified as SCC. The precision of this group of slides with respect to the total count of the recovered patches was 95%, resulting in a recall of 80% (F1=87%). False positives of 2% of the patches were equivocal for plasmacytoma and PNST (1% each). Patches ignored as MCT were reported as SCC and histiocytoma (9% each), followed by plasmacytoma (3%).

All slides were primarily diagnosed as MCT by (5.85/6 votes by the pathologists), with a confidence of 93%. Regarding this group, the second differential diagnosis was histiocytoma in 11/20 slides, with an average of 4.5/6 of votes and with a confidence of 5%. The third differential diagnosis was plasmacytoma on 11/20 slides, with a votes average of 4.5/6 and a confidence of 1%. This means that of the options provided, 4.5/6 votes agreed that the most likely differential diagnosis of MCT is histiocytoma and the third is plasmacytoma.

4.4.5 Histiocytoma

Out of the 20 slides, 19 were successfully classified (95% accuracy) and one was incorrectly classified as plasmacytoma. The precision was 80% with a recall of 81% (F1=81%). Confusion of the total number of patches that were erroneously classified as histiocytoma occurred mainly within MCT (8%) and plasmacytoma (5%) and among the ignored histiocytoma patches mainly within plasmacytoma (12%) and SCC (3%). The residual false negatives were less than 1% per remaining tumor.

All slides were primarily diagnosed as histiocytoma with 5.3/6 votes in average and with a confidence of 87%. From this group, the second diagnosis assigned by the pathologists (4.4/6 votes) was MCT, with a confidence of 1%; the second most voted was plasmacytoma for this category. The third diagnosis mentioned by pathologists (4.2/6 votes) was plasmacytoma, with a confidence of 2%; the second most voted in this category was MCT. This means that from the options provided, 4.4/6 votes agreed that the most likely differential diagnosis of histiocytoma is MCT and the third is plasmacytoma.

4.4.6 SCC

Within the 20 slides labeled as SCC, 18 were correctly classified (90% accuracy) and 2 slides (10%) were misclassified as plasmacytoma. The precision was 70% and recall was

85% (F1=77%). Matrix confusion shows that false positives were mainly classified within MCT (9%), PNST (7%) and plasmacytoma (6%), and that patches ignored as SCC were classified as PNST and histiocytoma in 3%, respectively. Confusion of patches that were ignored as SCC and classified as plasmacytoma was slightly more than 2%.

The six pathologists primarily diagnosed all slides as SCC (6/6), with a confidence of 95%. This shows the third most homogenous results. Trichoblastoma was the most voted secondary diagnosis in 19 of 20 slides, with an average of 5/6 of votes and a confidence of 3%. Melanoma was the most voted tertiary diagnosis in 19/20 slides, with an average of 4.3/6 votes and a confidence of 1%. This means that, from the options provided to pathologists, trichoblastoma is the main differential diagnosis of SCC and melanoma the third differential diagnosis.

4.4.7 Plasmacytoma

As in the previous case, 18 slides were correctly classified as plasmacytoma (90% accuracy). Of the remainder, one slide was incorrectly classified as melanoma and the other as SCC. This tumor type obtained a precision of 75% with respect to the total number of classified patches, with a recall of 80% (F1=77%). Of the remaining patches incorrectly classified as plasmacytoma (false positives), there was confusion mainly with histiocytoma in more than 10% of the patches, as well as PNST (4%) and MCT (3%). Of the retrieved patches that were ignored as plasmacytoma (false negatives), they were mainly classified as SCC (6%), histiocytoma and melanoma (5% each).

All pathologists correctly diagnosed 19/20 slides as plasmacytoma (5.7/6 votes in average), with a confidence of 82%. In one slide, 3/6 votes were for MCT as the primary diagnosis, with a confidence of 83%. From this slide, 3/6 votes were for plasmacytoma as second diagnosis, with a confidence of 26%. As differential (secondary) diagnosis, histiocytoma was the most voted tumor in 13 slides (3.6/5 votes of the pathologists in average, confidence of 15%), followed by MCT in 6 slides (3/6 votes in average, confidence of 21%). Finally, as a tertiary diagnosis, the most voted tumor was MCT on 15 slides (3.7/6 votes in average, confidence of 5%). From this group it can be said that from the choices given to the pathologists, the majority agree that the differential (secondary) diagnosis of plasmacytoma is histiocytoma, followed by MCT in the third differential diagnosis.

Table 5. Recall (sensitivity) and precision (positive predictive value) of pathologists' diagnoses and algorithm classification in dataset 2.

Tumor	Recall Pathologist consensus			Recall Algorithm
	Median	Min	Max	Estimate
Histiocytoma	0.925	0.750	1.000	0.800
MCT	0.975	0.950	1.000	0.900
Melanoma	0.950	0.900	0.950	0.950
Plasmacytoma	0.775	0.700	0.950	0.950
PNST	1.000	1.000	1.000	1.000
SCC	1.000	1.000	1.000	0.950
Trichoblastoma	1.000	0.950	1.000	1.000
Tumor	**Precision Pathologist consensus**			**Precision Algorithm**
	Median	**Min**	**Max**	**Estimate**
Histiocytoma	0.848	0.800	0.900	1.000
MCT	0.930	0.864	0.952	1.000
Melanoma	1.000	0.900	1.000	1.000
Plasmacytoma	0.944	0.762	1.000	0.760
PNST	0.952	0.952	1.000	0.952
SCC	0.952	0.952	1.000	0.950
Trichoblastoma	1.000	1.000	1.000	0.952

MCT: mast cell tumor; PNST: peripheral nerve sheath tumor; SCC: squamous cell carcinoma.

Table 6. Comparison of the combination of recall and precision (F1 score) with respect to the individual pathologists' diagnosis (first column), the pathologists' consensus (column two) and the algorithm classification (column 3).

Tumor	Pathologists single median (min- max)	Algorithm
Histiocytoma	0.897 (0.789-0.976)	0.889

MCT	0.950 (0.905-0.976)	0.947
Melanoma	0.974 (0.900-0.974)	0.974
Plasmacytoma	0.828 (0.780-0.950)	0.844
PNST	0.976 (0.976-1.000)	0.975
SCC	0.976 (0.976-1.000)	0.950
Trichoblastoma	1.000 (0.974-1.000)	0.975

5. Discussion

5.1 Algorithm performance

A reliable automated diagnostic work-up for seven of the most important/common canine skin tumors is shown in this study. This algorithm was developed with complete annotations on 350 WSI. Although we focused our annotations on all seven tumor types,we also created annotations for normal skin structures (Table 1). The major sub- compartments annotated in the skin were dermis, epidermis and subcutis (subcutaneousfatty tissue and muscle). Likewise, we created classes for tissue types that are not part ofthe skin per se, but were present in some cases such as ear cartilage, bone in the paws (mainly melanoma and SCC) (Marinoet al., 1995) or inflammation and necrosis as a secondary response to tumor growth. Artifacts were ignored and excluded in all slides.

In many cases, the stroma was present in high amounts (e.g. PNST or trichoblastoma), so the stroma was annotated as part of the tumor. In tumors where it was difficult to distinguish the exact borders (e.g. in round cell tumors or SCC), the borders were definedas the region where normal tissue began/ended (Figure 1J-L). In order to obtain accurateresults, the annotations were systematically made with the polygon tool (Figure 1)(Aubreville et al., 2018). We performed the skin annotations starting from the external surface (epidermis) towards the deep surface (subcutaneous), usually from left to right (Figure 1F), similar to the system described by Lindman et al. (Lindman et al., 2019) The delineation between the epidermis and the rest of the tissues was the simplest but still time-consuming, due to de great amount of it in most of the slides (Table 1, Figure 1D-E). The separation of the tumor and dermis was the most laborious and time-consuming because all tumors had at least a small dermal involvementand because we wanted to avoid overlapping annotations on the entire slide. In those slides where the tumors were easier to delineate, we annotated the dermis after epidermis; in those where there was no clear division between dermis-tumor-subcutis, weannotated the tumor after the epidermis and the rest after. No annotation was overlapped.

Specifically, in SCC, tumor delineation was complicated by the high frequency of inflammation (Figure 1J-L), desmoplasia and necrosis. Inflammation in SCC is one of the most notorious and common features observed although the role of its presence is controversial (Cerezo-Echevarria et al., 2020; Santana et al 2016). Santana et al demonstrated that the presence of inflammation in SCC, regardless of the degree of differentiation, is common and is usually characterized by macrophages, lymphocytes and plasma cells (Santana et al., 2016). For this reason, and due to its high difficulty during tumor delineation, we decided to include subacute-chronic inflammation secondary to the tumor in most cases. Acute inflammation (mostly necrotic and neutrophilic) was annotated as part of the inflammation/necrosis class only in those cases where there was ulcerationor external inflammation; when the inflammation involved inner tumoral structures, it was annotated as part of the tumor (Figures 1A-C, 1G-L). Our results showed that within the CNN during testing, the algorithm was able to correctly diagnose 90% of the SCC slides; however, confusion with plasmacytoma resulted in 10% of the slides labeled as SCC (Figure 5). Likewise, in the slides labeled as MCT and plasmacytoma, the reverse occurred, as 5% of the slides of these tumors were incorrectly classified as SCC (Figure 7). We speculate that the confusion of patches within our database in this group was a consequence of the complexity of the tumor (high frequency of subacute-chronic inflammation among neoplastic groups of cells), as well as the objectivity of the algorithm. The evaluated patches demonstrate that the algorithm had the greatest difficulty of classification in those areas with the highest number of inflammatory cells (Figure 5) and classified them with respect to most of the cells present (plasma cells, histiocytes or lymphocytes). It could be that in some patches, the differentiation between round cells caused conflicts because although no or only few mast cells were observed as part of the inflammation, the algorithm decided to classify some patches as MCT. In the case that was misclassified as plasmacytoma, there was a large amount of inflammation and little squamous cell tumor density. We believe this is justified during the annotation process. Nevertheless, our data are of high relevance since to date, no algorithm has been developed that has achieved this using H&E-stained WSI in cutaneous SCC.

Following the issue of inflammation and necrosis, as mentioned above, in addition to the normal skin structures and the seven tumor types, we decided to create an annotation class for inflammation and necrosis in order to achieve a more adequate segmentation during the testing of the algorithm. In total, 719 inflammation/necrosis annotations were achieved, with an area of 2050.16 mm^2 (Table 1) and successful segmentation (Figure 3).Although it was not our goal to determine and quantify necrosis within tumors, this is an important finding in the diagnostic and prognostic understanding of specific situations. Theimportance of necrosis within tumors has been proven (Hanahan & Weinberg, 2011), as determining its presence and extent can provide useful information for the diagnosis andprognosis of some tumors such as soft tissue sarcomas, like PNST (Kuntz et al., 1997) and melanoma (Smith et al., 2002). In these two tumors in our database, the delineation of necrosis was laborious because it was randomly distributed inside and outside the tumor (Figure 1A-C, 1G-H); however, segmentation was successfully achieved and the necrotic regions were excluded before classifying the tumor type by the algorithm (Table 2, Figure 3). A good performance of ML and DL in DP has been reported during the evaluation of tumor-related necrosis mainly in osteosarcoma (Arunachalam et al., 2019; Fu et al., 2020; Ho et al., 2020), due to its great usefulness during patient assessment in human chemotherapy (Kang et al., 2017).Likewise, Arunachalam et al. (Arunachalam et al., 2019) described a reliable model for identification and quantification of necrosis within osteosarcoma with potential for use in other tumor types. In the same way, we believe that our method and model could be implemented in veterinary pathology in tumors of dogs and most probably other species.

Our model classified all trichoblastoma slides correctly, with excellent performance as its precision and recall were 94% each (F1=94%). These results are the highest and most homogeneous of the seven tumors we included in our database, training and testing. As mentioned above, the annotations of this tumor were the most easily made because of its clear demarcation from the rest of the normal histologic structures of the skin and becauseit does not tend to be contiguous with the epidermis (Figure 1D-E). We suspect that these characteristics, in addition to the tissue morphology of this neoplasm, were fundamental factors for the excellent performance of our model. Trichoblastoma is the most common cutaneous follicular tumor in dogs (Abramo et al., 1999; Goldschmidt et al., 2018;

Goldschmidt, 1998) and its histologic features differentiate it from most cutaneous tumors and tumor-like lesions (Abramo et al., 1999; Goldschmidt et al., 2018; Goldschmidt, 1998; Wiener, 2021). Despite its different histologic subtypes (ribbon, medusoid, trabecular, granular, and fusiform), its histologic features are unique in our database. This probably also explains the efficiency of our model. It would be interesting in future research to test whether a differentiation between trichoblastoma and its differential diagnoses (trichoepithelioma, tricholemmoma, basal cell carcinoma, etc.) can be created through ML and DL in WSI. Following the previous line of discussion, within the confusion matrix it can be determined that the algorithm had a slight difficulty indifferentiating it with SCC, as 2% of the evaluated patches of this group were classified as such (Table 4). This finding is interesting as it means that the algorithm was able to identify similarities between tumors of follicular origin. This is the first study in veterinary medicine to include trichoblastoma within a CNN training in WSI of canine tumors.

All slides labeled as PNST were correctly classified (100% accuracy), with a precision of 91% and a recall of 86% (F1=88%). Of the total number of patches recovered during the test in this group of tumors, 3% were classified as SCC and 2% as melanoma; likewise, the tumor with the highest number of false negatives was SCC (Table 4). Previously we discussed one of the main characteristics of SCC that caused confusion in our model, such as inflammation and necrosis; however, within its histological features, the high desmoplastic activity is well known (Goldschmidt et al., 2018; Goldschmidt, 1998; Zainab et al., 2019) which, as in inflammation, the relevance of its presentation is not yet fully elucidated and is still under discussion (Zainabet al., 2019).

Likewise, one of the main components of PNST (and soft tissue sarcomas) in dogs is the presence of variable amounts of fibrovascular stroma (Dennis et al., 2011; Hendrick, 1998), very similar to that present in SCC. In reviewing the patches, we found that our model had difficulty differentiating between the two tumor types in a few of them. Since stroma is a secondary reaction to the presence of both tumors, we determined that it is very important to include any type of stroma within the annotations of tumors with such ability, because at the end of the day, the difficulties in its classification were not relevant for the correct classification and excellent performance of our algorithm. Our method of annotation and training resembles that of Foersch et al, who developed an algorithm for the identification and differentiation of soft tissue sarcomas in humans with a high accuracy, similar to ours (Foersch et al., 2021). Although we did not replicate the previously mentioned method, we can conclude that in the same way, our model can assist pathologists, shorten diagnostic intervals and increase their accuracy and confidence, regardless of the degree of expertise. Although there are studies of AI in softtissue sarcomas in dogs (mainly in diagnostic imaging such as tomography) (Ye et al., 2021), at present date its application in WSI has been scarcely investigated. This is one of the first studies to include and investigate the performance of AI in a type of soft tissue sarcoma in dogs with complete annotation in WSI. The cellular characteristics and their relevance within ML and DL in WSI will be addressed in the nextparagraph.

The performance of our model during the classification of the slides labeled as melanoma reached 95% accuracy (19/20), with a precision of 91% and a recall of 90% (F1=90%). In fact, this number is surprising since it was hypothesized that this tumor would cause problems during testing due to its complexity and variability in pigmentation and histological patterns and cell forms (pleomorphism), to such an extent that even for pathologists its correct diagnosis can be challenging without using special tests (for example IHC) (Goldschmidt, 1998; Smedley et al., 2011). Most likely, we believe the amount of pigment greatly facilitated itsclassification, since from the slides we chose, 45% were heavily pigmented melanomas, 35% had less than 50% pigment and 20% were amelanotic melanomas. One of these slides was incorrectly classified as PNST (Figure 6). This case is very particular and interesting within our database, since very similar results occurred in the human vs. machine challenge. The aforementioned slide corresponds to a melanoma whose histological characteristics contained a large number of spindle cells and little melanin pigment (amelanotic spindle cell melanoma). As expected, the algorithm had great difficulty classifying this case and the performance was not as precise as it determined that 44% of the patches corresponded to PNST, 35% to trichoblastoma, and only 13% to melanoma. Spindle melanomas are among the different subtypes that have been recognized (along with epithelioid melanoma), with absent or present pigment (Smedley et al., 2011). These melanomas are arranged in streams and interweaving bundles, very similar to soft tissue sarcomas (PNST or fibrosarcoma) but with nuclear pleomorphism and karyomegaly, so those amelanotic spindle melanomas are often a diagnostic challenge without the utilization of specific tests such as IHC (Goldschmidt, 1998; Ramos-Vara & Miller, 2011; Smedley et al., 2011). Outside this specific case, our algorithm had an accuracy similar to that of the 6 pathologists who evaluated the same cases, so once again the model we developed has great diagnostic utility and could even be established as an additional tool to software-assisted decision making in diagnostic and research laboratories in veterinary pathology. It is worth mentioning that the usefulness of the development of CNN in DP for the diagnosis of melanoma has already been proven on several occasions (Norgan et al., 2018; Wang et al., 2020); however, in veterinary pathology it is still a little explored area.

It would be worthwhile to investigate in depth its vast possibilities and future applications.

5.2 Algorithm performance: round cell tumors

For the round cell tumors included in our database, histiocytoma and MCT had a very good classification accuracy of 95% (19/20) each and plasmacytoma 90% (18/20). Although MCT and histiocytoma had the same number of correctly classified slides, the precision of the individual patches was different, with MCT having an excellent precision (95%) and histiocytoma a good precision (80%). However, the recall of both and plasmacytoma was similar (MCT=80%, histiocytoma=83%, plasmacytoma=80%). Likewise, the recall of these three tumors was the lowest in our results (less than 85%). This means that out of the seven tumors, our model showed greater difficulties in collecting the positive patches compared to the rest of the tumors, which obtained a recall rate equal to or greater than 85%. We acknowledge that there is considerable discussion among researchers concerning the facility for a pathologist to differentiate round cell tumors from each other, which in addition to those we investigated, cutaneous lymphoma, amelanotic melanoma, neuroendocrine tumors and transmissible venereal tumor (TVT) are mentioned (Cangul, 2001; Meuten, 2016). It is also well known that due to their similar cellular appearance and tissue arrangement, it is a great challenge for pathologists to differentiate them without the utilization of special stains or complementary tests such as IHC, especially in poorly differentiated tumors (Sandusky et al., 1987).

Among these three tumors, plasmacytoma has a special complexity. In most of the plasmocytomas, regardless of the degree of differentiation, it is common to observe atypical morphologic configurations like binucleations, multinucleations, variable amount of amyloid, nuclear atypia and cellular pleomorphism (Baer et al., 1989; Banerjee et al., 2004). Likewise, the cellular arrangement, which is usually sheet-like, atypical presentations in the form of pseudoglandules have been reported (McHale et al., 2018). We believe that the previously mentioned features mainly influenced the results of our model, as this was the tumor with the lowest precisionand recall (F1=77%). Although this tumor typically offers no diagnostic challenge for the pathologist, some subsets of it might cause difficulties in differentiating it from other neoplasms such as amelanotic melanoma. Although little research has been done on DLand CP with a focus on round cell tumors in canines, Salvi et al. developed an algorithm very similar to the one we describe with images collected from WSI (Salvi et al., 2021). However, comparing their results with ours, and since they were very similar, we can conclude that our algorithm may be reproducible and of high benefit in decreasing error rates during round cell diagnosis, not only in dogs but also in other species. Likewise, although our intention was not to stage MCT in their grades of malignancy as Salvi et al. did, we do not exclude the possibility that in future research, an algorithm for the automated diagnosis of canine cutaneous tumors can be created that can also achieve acorrect staging of malignancy. Our findings are of vital importance, since a correct diagnosis of round cell tumors in dogs is important to determine the prognosis and treatment of patients and this model could potentially be implemented as a diagnostic support tool for the daily workflow in pathology laboratories with equally efficient results but in less time and with lower cost than immunomolecular tests.

Salvi et al. encountered similar difficulties in developing an algorithm for automated detection of round cell tumors. They concluded that the amount of inflammatory cells present in certain tumors (e.g. histiocytoma or SCC) might influence the decisions made by the algorithm (Salvi et al., 2021). In the same way, our results can be compared with those mentioned above, as with respect to our results, we believe that it is indispensable to take these features into account in future attempts to improve the automated classification of round cell tumors in dogs.

A well-illustrated example of the distribution of patches in a WSI during the algorithm classification is shown in Figure 8, where the histiocytoma was incorrectly classified as a plasmacytoma. Another interesting finding in our model is the automatic identification of the epidermal reaction secondary to tumor growth in some patches and its classification as SCC (Figure 4).

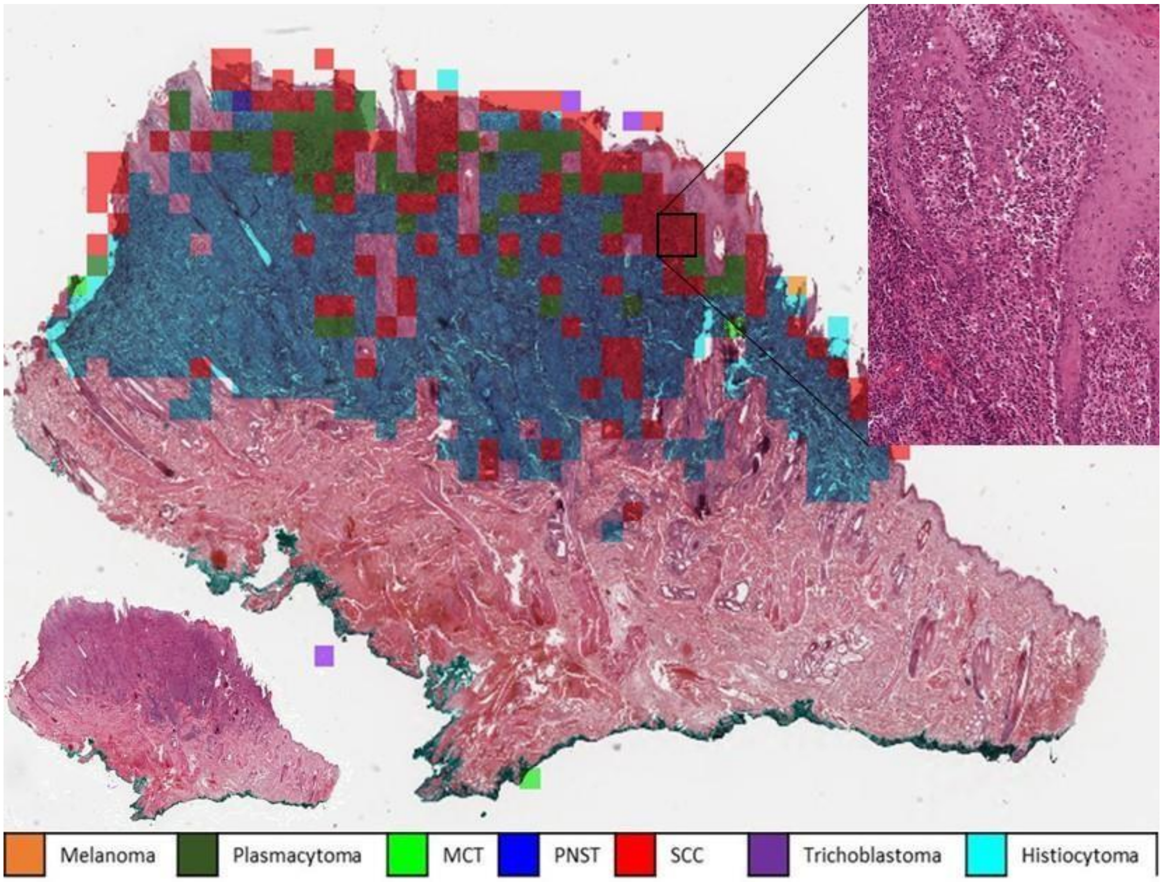


Figure 4. Histiocytoma. Automatic classification algorithm in WSI (normal histology on the lower left). Note that most of the patches chosen by the algorithm that were correctly classified as histiocytoma are situated in the central and lower region of the tumor (blue patches) and in the upper region, close to the epidermis, most of the patches were classified as SCC (red patches). Of the total number of patches in this WSI (n=495), 66% were correctly classified as histiocytoma and 22% as SCC. Additionally, 7% of the patches were classified as plasmacytoma, also distributed on the superficial surface of the tumor. On the upper right is a magnification of a region that was classified as SCC; the epidermis is identified with an irregular, pseudocarcinomatous acanthosis, subacute inflammation and neoplastic cells.

WSI, H&E. MCT: mast cell tumor; PNST: peripheral nerve sheath tumor; SCC: squamous cell carcinoma.

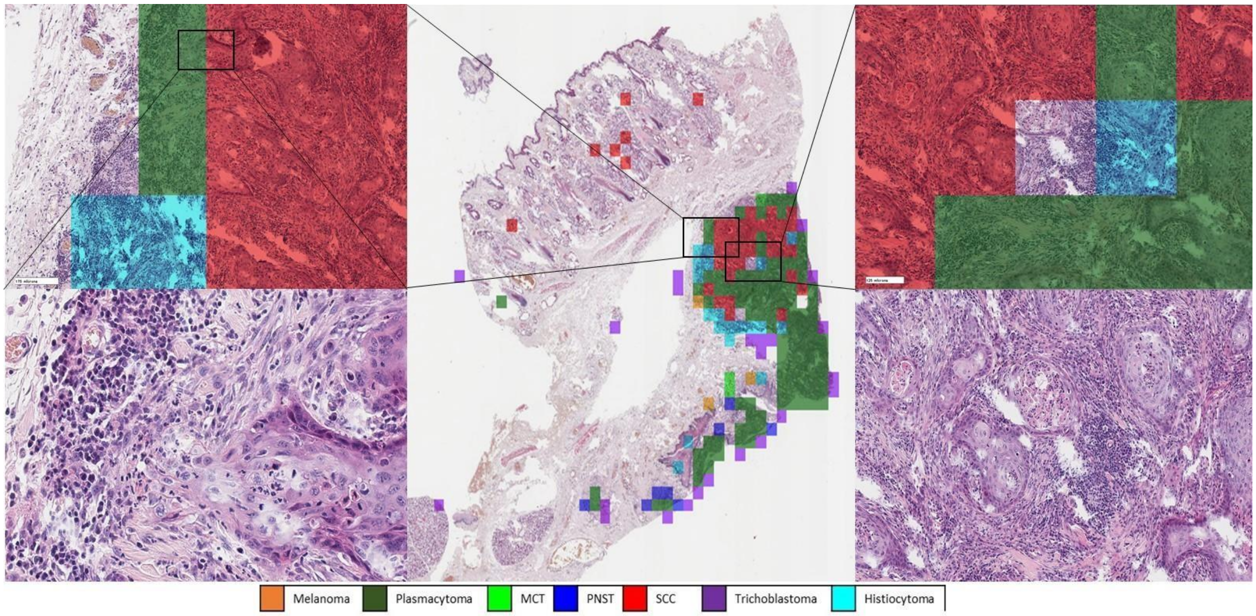


Figure 5. SCC misclassified as plasmacytoma. Out of the total number of patches recovered during segmentation (n=204), 51% were classified as plasmacytoma (green patches) and 21% as SCC (red patches). In the left magnification, a transition zone is observed; note that the neoplastic epithelial cell clusters (red) are properly classified and the patches from that region arranged in the periphery correspond to subacute inflammation which was classified as plasmacytoma (green) and histiocytoma (blue). In the same way, in the magnification on the right side, the neoplastic groups in red (SCC) and in the area with predominant inflammation classified as plasmacytoma and 1 patch as histiocytoma.

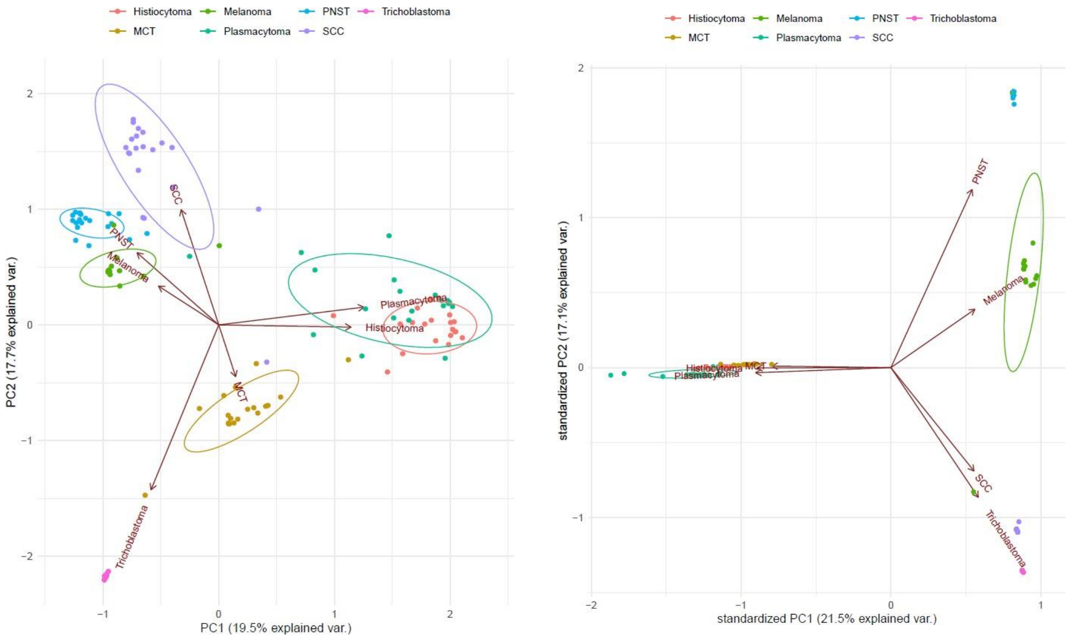
WSI, H&E. MCT: mast cell tumor; PNST: peripheral nerve sheath tumor; SCC: squamous cell carcinoma.

5.3 Comparison of algorithm performance against human performance

In order to validate the results of the algorithm and compare its results with the expertise of the pathologists, a human vs. machine challenge was performed in which 6 experienced and board-certified pathologists were provided with the same 140 slides used in the test set of the algorithm. The qualitatively accuracy of the pathologists by diagnosing the slides was 98%, i.e. of the total number of slides (n=140), 137 were correctly diagnosed. The group of slides where misclassification or mismatch was observed was in the slides labeled as melanoma (2/20) and plasmacytoma (1/20).

Regarding slide 98, whose diagnosis corresponds to melanoma, in the same way as the algorithm, 4/6 pathologists diagnosed it as PNST and 2/6 correctly as melanoma. This case was previously discussed. However, it is very interesting that the algorithm behaved similarly to the majority of pathologists (Table 6). It is worth remembering that in this experiment, the pathologists were provided only with WSIs without the possibility of performing special stains or complementary tests (e.g., IHC), so they had to depend on their first guess (visual skills). In addition, the difficulty that can arise when diagnosing melanomas with H&E alone, especially in poorly pigmented tumors, has been described on numerous occasions (de Wit et al., 2004; Jungbluth, 2008; Koenig et al., 2001; Ohsie et al., 2008), and therefore other tools are usually necessary. As already mentioned, this case represented a melanoma with apredominance of spindle cells and scarce amount of pigment with strong positivity in IHCagainst Melan-A antibodies (Figure 6). With respect to the majority vote and the algorithmdecision, we could conclude that our results serve as a basis for determining the inclusionof AI to support pathologists in the routine workflow. This provides a good starting point for discussion and further research. We believe that future research should be devoted tothe development of algorithms that can reach the value of a molecular complementary test.

The other case in the melanoma group that was incorrectly diagnosed by the majority of pathologists (4/6, precision 91%) was diagnosed as SCC. Two of the six pathologists correctly diagnosed this case. This case corresponds to an amelanotic melanoma composed of mostly epithelioid cells and with clustered and nested arrangement, as described in the literature (Goldschmidt, 1985; Smith et al., 2002). The behavior ofthe algorithm was distinctive, since out of the 100% of the patches evaluated by the algorithm, 50% of them were correctly classified as melanoma, obtaining the majority of the votes, followed by trichoblastoma (26%) and SCC (18%). Although the algorithm classified it correctly, the tumors in our database with which it found the most confusion were trichoblastoma and SCC, similar to the pathologists. This can only be explained by taking into account that the pathologists performed the diagnosis at a slide level, while the algorithm performed it at a patch level, so the pathologists' diagnostic confidence, without complementary diagnosis techniques has to be arbitrary while the algorithm's decision-making is more accountable (honest?) (Graphic 1).



Graph 1. A. Distribution of algorithm classification regarding the tumor type. B. Distribution of pathologists' consensus with respect to definitive diagnosis. Each point represents a WSI that was grouped into a class with its precision of occurrence. n=140. SCC: squamous cell carcinoma; MCT: mast cell tumor; PNST: peripheral nerve sheath tumor.

Another explanation is that diagnosis among different pathologists may become subjective and subject to considerable inter-observer variability (Bueno-de-Mesquita et al., 2010; Orlando et al., 2016; Schnitt, 2001). However, in recent studies, Bertram et al. developed an algorithm to support the detection of mitoses in WSIs of canine MCT by comparing it with the visual identification of board certified pathologists, demonstrating that the assistance of algorithms in routine diagnosis is of high utility when there is high inter-observer variability (Bertram et al., 2021; Bertram et al., 2020). Likewise, we propose that our model can be implemented as part of a tool to support pathologists in daily workflow, especially in dermatologic oncology.

In the last slide where there was an incorrect diagnosis, which was a plasmacytoma (CD79 positive) (Ramos-Vara et al., 2007), 3/6 pathologists diagnosed it as MCT. However, the average confidence with which the diagnosis was made was 83% and as a differential diagnosis 3/6 pathologists considered plasmacytoma with an average confidence rate of 26%. Alike, our model misclassified this case, as of the total number of patches recovered, 44% were recognized as SCC, 17% as melanoma and 12% as histiocytoma. Plasmacytoma obtained only 8% precision (Figure 7). We previously discussed the frequent difficulty encountered during the diagnosis of round cells, especially when the degree of differentiation is very low. By comparing the results of Salvi et al. and their algorithm for the automated detection of round cell tumors (Salvi et al., 2021), we consider that in the same way our model could increase the efficiency of pathological diagnosis, as well as serve as a second opinion tool for pathologists. Furthermore, this study is a milestone for automated classification of round cell tumors in WSIs using CNN.

Tables 5-6 show a summary comparison of the combination of pathologists' precision, recall and F1 score with respect to those of our model and in the graphic 2 an visual overview of recall, precision and F1 score of both, the algorithm and the pathologists consensus. It is important to note that the behavior of our model is very similar to that of the pathologists only when individual values are considered; we believe this is due to frequent inter-observer variability. However, when compared to the pathologists' consensus, the precision and recall clearly increases. The graphic 1, nevertheless, shows the simplicity with which each pathologist diagnosed the 140 slides compared to the different possibilities that the algorithm defined for each slide (slide level vs patch level). This is clearly a consequence of the fashion of human reasoning versus the prediction of an AI algorithm, since it can be said that the pathologists determined their diagnosis at a slide level, considering general and particular, tissue and cellular features as a whole to determine the ultimate decision, with respect to their expertise. On the other hand, the algorithm segmented the slide, determined the tumor area and divided it into hundreds of patches, analyzing it in detail one by one and determining an absolute decision for each of them (Graphic 1). Let us remember that in the end, the class (type of tumor) that is counted with the highest number determines the final decision of the algorithm. Since the sensitivity and positive predictive value behaved similarly to that of the six pathologists, we can conclude that our model could function as a diagnostic support tool in a similar way to that which could support other types of tests, molecular for example.

Nevertheless, this assumption might be addressed in future studies.

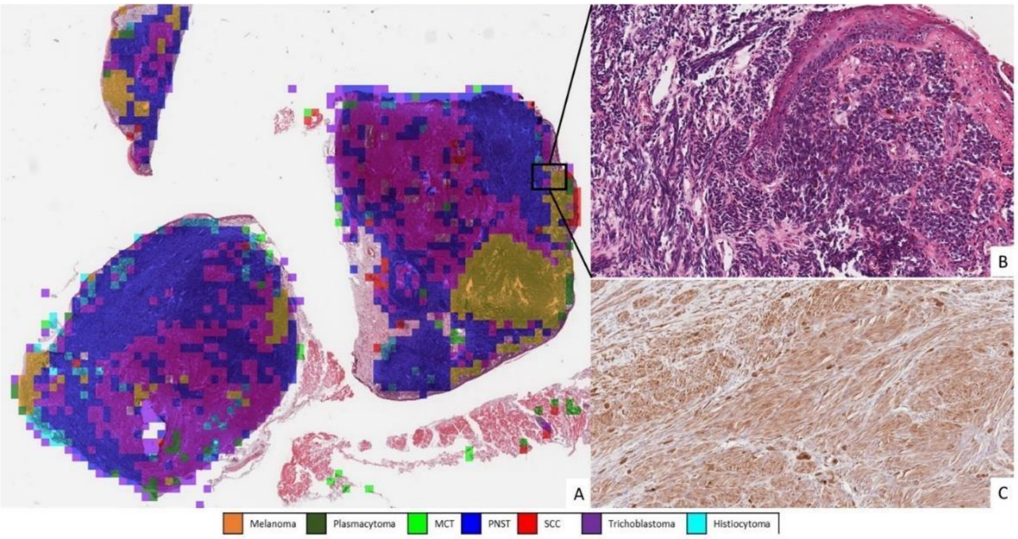


Figure 6. Amelanotic melanoma incorrectly classified as PNST. A. Out of the total number of patches (n=1976), 44% were classified as PNST (blue), 35% as trichoblastoma (purple) and a mere 13% as melanoma (orange). B. Magnification of a transitional region between patches classified as melanoma (right side) with a predominance of subepidermic epithelioid cell nests and patches classified as PNST (left side) with a predominance of spindle cells. WSI, H&E. C. Melan-A IHC with strong cytoplasmic positivity. MCT: mast cell tumor; PNST: peripheral nerve sheath tumor; SCC: squamous cell carcinoma**.**

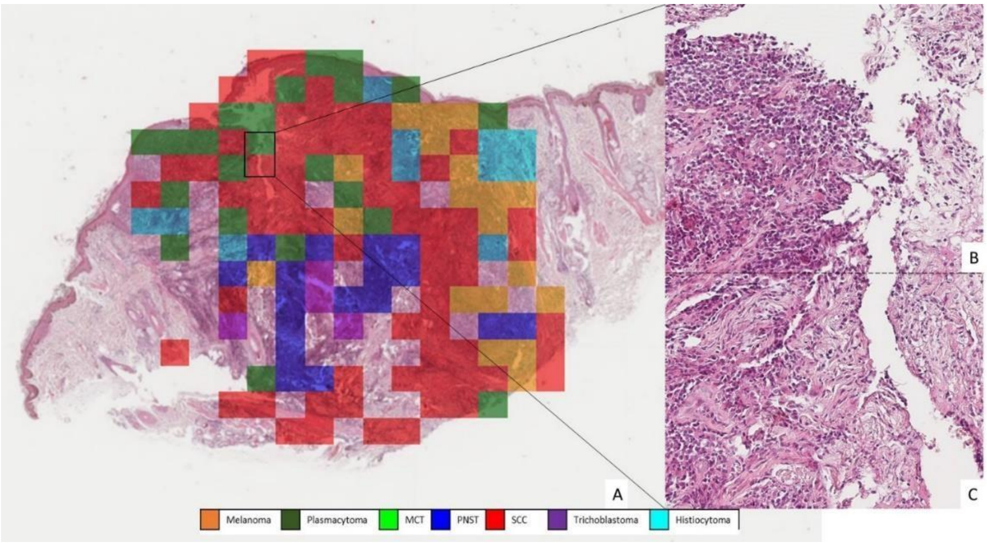


Figure 7. Plasmacytoma misclassified as SCC. A. Out of the total number of patches recovered during segmentation, 44% were classified as SCC (red), 12% as histiocytoma and only 8% as plasmacytoma. B. Magnification of a region classified as plasmacytoma (green) with a predominance of neoplastic plasma cells. C. Patch classified as SCC with low density of neoplastic cells and high density of stroma. WSI, H&E. MCT: mast cell tumor; PNST: peripheral nerve sheath tumor; SCC: squamous cell carcinoma.

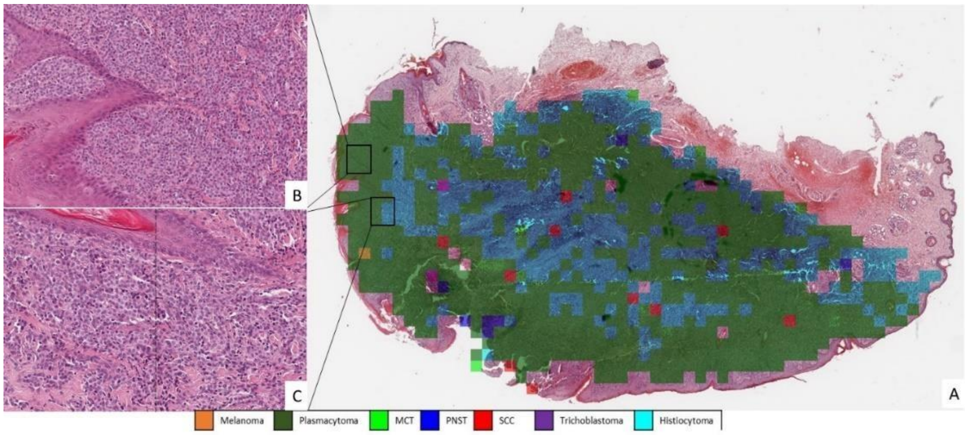
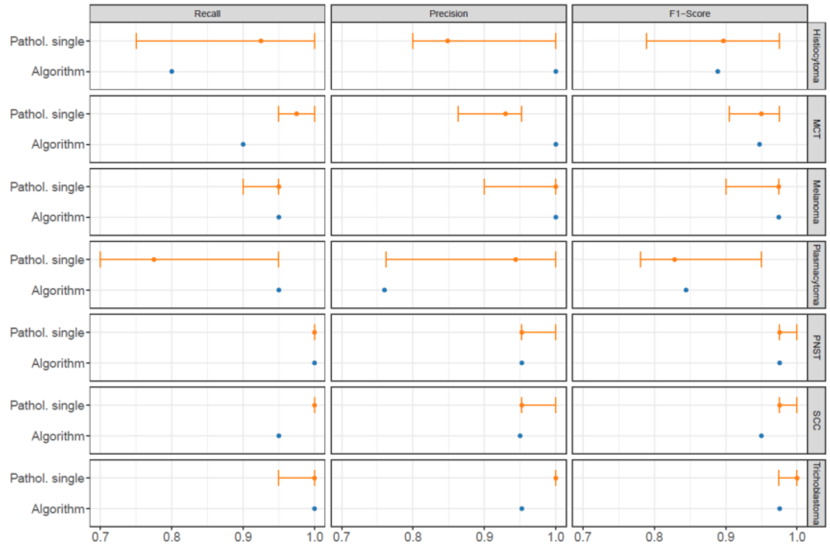


Figure 8. Histiocytoma misclassified as plasmacytoma. A. Out of the total number of patches retrieved during segmentation (n=897), 64% were classified as plasmacytoma (green) and 32% as histiocytoma. Note the distribution of correctly classified patches in the center and at the bottom margin of the tumor. B. Magnification to an area completely classified as plasmacytoma.C. Transition zone between plasmacytoma (left) and histiocytoma (right). WSI, H&E. MCT: mast cell tumor; PNST: peripheral nerve sheath tumor; SCC: squamous cell carcinoma.



Graph 2. Overview of recall (sensitivity), precision (positive predictive value) and the combination (F1 score) of the individual diagnoses of the pathologists, consensus and algorithm. KI. Algorithm.

6. Conclusions / Summary

- ✓ Anatomic pathology is a medical specialty with a basic doctrinal body that makes it, on the one hand, an autonomous academic discipline and, on the other hand, a functional unit in medical care as a diagnostic tool. The visualization and interpretation of H&E-stained slides remains the basis of pathological analysis and diagnostic medicine for more than a century.

- ✓ DP is an emerging technology in pathology, in which a scanner converts glass slides into WSI that can be viewed, analyzed and managed on a screen with the help of visualization software, after which they are stored digitally.

- ✓ It can facilitate faster and more efficient diagnoses and prognoses, more flexible collaboration and high-quality case documentation. Online storage of digital slides also opens the door to DL and AI applications.

- ✓ The pathologist requires extensive and constant training, usually based on following algorithmic decision trees that bring together a large amount of information and its association with respect to cellular and tissue structures (visual skills) in order to describe lesions and determine an appropriate diagnosis. However, image interpretation is not always consistent among pathologists. Emerging AI technologies, specifically ML and DL are now a state-of-the-art tool that is routinely used in human pathology and in many veterinary pathological diagnostic institutions. They have also proven to be very useful in reducing disagreements between observers during H&E slide interpretation.

- ✓ Our results highlight the similarities that artificial intelligence and human intelligence share with respect to histopathologic diagnosis; however, this is more evident in certain types of tumors, mainly round cell tumors. For example, our model presented greater difficulty in differentiating round cell tumors from SCC and the pathologists' consensus presented greater difficulty in differentiating round cell tumors from each other.

✓ Our results also highlight the feasibility of including artificial intelligence as a support tool in diagnostic and research oncologic pathology with future applications in other species and other tumor types.

7. Zusammenfassung

„Automatisierte Diagnose von sieben wichtigen Hauttumoren bei Hunden mit einem neuronalen Faltungs-Netzwerk (CNN) auf H&E-gefärbten Ganzpräparatbildern (WSI)"

Die mikroskopische Untersuchung von HE-gefärbten Objektträgern ist der Goldstandard für eine Vielzahl von Krankheiten. Speziell in der Onkologie ist sie nicht nur für eine präzise Diagnose, sondern auch für das Staging von Tumoren und die Evaluierung ihrer Grenzen entscheidend. In den letzten Jahrzehnten, mit dem Aufkommen der Digitalen Pathologie (DP) und der Whole Slide Images (WSIs), steht die Image-Analyse und die Entwicklung von Algorithmen zur Durchführung spezifischer Aufgaben auf WSIs an vorderster Front der Forschung in der Pathologie, mit überwältigenden Ergebnissen. In dieser Studie beschreiben wir einen funktionellen Algorithmus zur automatischen Erkennung von sieben großen Hauttumoren bei Hunden: Trichoblastom, Plattenepithelkarzinom (SCC), peripherer Nervenscheidentumor (PNST), Melanom, Histiozytom, Mastzelltumor (MCT) und Plasmozytom. Wir haben 350 H&E-gefärbte Objektträger (70 pro Tumorart) ausgewählt, digitalisiert und mit Anmerkungen versehen, um eine Datenbank zu erstellen, die in Trainings- (n=245 WSIs), Validierungs- (n=35 WSIs) und Testdaten (n=70 WSIs) unterteilt ist. Anschließend wurde ein neuronales Faltungsnetzwerk (CNN) entwickelt und die Effizienz des Algorithmus an 140 neuen WSIs (20 pro Tumorart) getestet. Die Klassifizierungsgenauigkeit auf Objektträgerebene erreichte 95 % (133/140 WSIs), die Präzision auf Patch-Ebene lag bei 85 %. Dieselben 140 WSIs wurden sechs zertifizierten Pathologen zur Diagnose vorgelegt, die eine ähnliche Genauigkeit auf Objektträgerebene von 98 % erreichten (137/140 WSIs). Unsere Ergebnisse zeigen, dass der Einsatz von künstlicher Intelligenz als Hilfsmittel in der diagnostischen und forschenden onkologischen Pathologie machbar ist und in Zukunft auch bei anderen Spezies und anderen Tumorarten angewendet werden kann.

8. Literature

Abels, E., Pantanowitz, L., Aeffner, F., Zarella, M. D., van der Laak, J., Bui, M. M., Vemuri, V. NP., Parwani, A. V., Gibbs, J., Agosto-Arroyo, E., Beck A. H & Kozlowski, C. (2019). Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. *Journal of Pathology, 249*(3),286-294. doi:10.1002/path.5331

Abramo, F., Pratesi, F., Cantile, C., Sozzi, S., & Poli, A. (1999). Survey of canine and feline follicular tumours and tumour-like lesions in central Italy. *Journal of Small Animal Practice, 40*(10), 479-481. doi:DOI 10.1111/j.1748-5827.1999.tb02999.x

Aeffner, F., Wilson, K., Martin, N. T., Black, J. C., Hendriks, C. L. L., Bolon, B., Young, G. D. (2017). The gold standard paradox in digital image analysis: manual versus automated scoring as ground truth. *Archives of Pathology & Laboratory Medicine, 141*(9), 1267-1275. doi:10.5858/arpa.2016-0386-RA

Afework, A., Beynon, M. D., Bustamante, F., Cho, S., Demarzo, A., Ferreira, R., Tsang, H. (1998). Digital dynamic telepathology--the Virtual Microscope. *Proc AMIA Symp*, 912-916.

Alfaro Ferreres, L. (2001). *Manual de telepatología*. Pamplona: Club de informatica aplicada de la Sociedad Espanola de Anatomia Patologica.

Ang, J. C., Mirzal, A., Haron, H., & Hamed, H. N. A. (2015). Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM transactions on computational biology and bioinformatics, 13*(5), 971-989.

Aprupe, L., Litjens, G., Brinker, T. J., van der Laak, J., & Grabe, N. (2019). Robust and accurate quantification of biomarkers of immune cells in lung cancer micro-environment using deep convolutional neural networks. *PeerJ, 7*, e6335.

Araujo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., Campilho, A. (2017). Classification of breast cancer histology images using Convolutional Neural Networks. *Plos One, 12*(6). doi:ARTN e0177544

Arunachalam, H. B., Mishra, R., Daescu, O., Cederberg, K., Rakheja, D., Sengupta, A., Leavey, P. (2019). Viable and necrotic tumor assessment from whole slide images of osteosarcoma using machine-learning and deep-learning models. *Plos One, 14*(4). doi:ARTN e0210706

Assawawongkasem, N., Techangamsuwan, S., Piyaviriyakul, P., Puchadapirom, P., & Sailasuta, A. (2020). Involucrin, cytokeratin 10 and Ki67 expression in a three-dimensional cultured canine keratinocyte cell line in comparison to canine skin and cutaneous squamous cell carcinoma and a pilot study on custom-designed siRNA-INV transfection. *Thai Journal of Veterinary Medicine, 50*(1), 43-55.

Aubreville, M., Bertram, C., Klopfleisch, R., & Maier, A. (2018). SlideRunner, pp. 309-314. Springer Berlin Heidelberg, Berlin, Heidelberg, 2018.

Awan, R., Sirinukunwattana, K., Epstein, D., Jefferyes, S., Qidwai, U., Aftab, Z., Mujeeb, I., Snead, D. & Rajpoot, N. (2017). Glandular morphometrics for objective grading of colorectal adenocarcinoma histology images. *Scientific Reports, 7*(1), 1-12.

Baer, K. E., Patnaik, A. K., Gilbertson, S. R., & Hurvitz, A. I. (1989). Cutaneous plasmacytomas in dogs: a morphologic and immunohistochemical study. *Vet Pathol, 26*(3), 216-221. doi:10.1177/030098588902600305

Baines, S. J., Mcinnes, E. F., & McConnell, I. (2008). E-cadherin expression in canine cutaneous histiocytomas. *Veterinary Record, 162*(16), 509-513. doi:DOI 10.1136/vr.162.16.509

Banerjee, S. S., Verma, S., & Shanks, J. H. (2004). Morphological variants of plasma cell tumours. *Histopathology, 44*(1), 2-8. doi:DOI 10.1111/j.1365-2559.2004.01763.x

Bankhead, P., Loughrey, M. B., Fernández, J. A., Dombrowski, Y., McArt, D. G., Dunne, P. D., McQuaid, S., Gray, R. T., Murray, L. J., Coleman, H. G., James, J. A., Salto-Tellez, M. & Hamilton P. W. (2017). QuPath: Open source software for digital pathology image analysis. *Scientific Reports, 7*(1), 1-7.

Becker, A. (2019). Artificial intelligence in medicine: What is it doing for us today? *Health Policy and Technology, 8*(2), 198-205.

Bejnordi, B. E., Zuidhof, G., Balkenhol, M., Hermsen, M., Bult, P., van Ginneken, B., Karssemeijer, N., Litjens, G. & van der Laak, J.(2017). Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images. *Journal of Medical Imaging, 4*(4), 044504.

Bertram, C. A., Aubreville, M., Donovan, T. A., Bartel, A., Wilm, F., Marzahl, C., Assenmacher, CA., Becker, K., Bennett, M., Corner, S., Cossic, B., Denk, D., Dettwiler, M., Garcia-Gonzalez, B., Gurtner, C., Haverkamp, A. K., Heier, A., Lehmbecker, A., Merz, S., Noland, E. L., Plog, S., Schmidt, A., Sebastian, F., Sledge D. G., Smedley, R. C., Tecilla, M., Thaiwong, T., Fuchs-Baumgartinger, A., Meuten D. J., Breininger, K., Kiupel, M., Maier, A & Klopfleisch, R. (2021).Computer-assisted mitotic count using a deep learning-based algorithm improves interobserverreproducibility and accuracy. *Veterinary Pathology*. doi:Artn 03009858211067478

Bertram, C. A., Aubreville, M., Gurtner, C., Bartel, A., Corner, S. M., Dettwiler, M., Kershaw, O., Noland, E. L., Schmidt, A., Sledge, D. G., Smedley, R. C., Thaiwong, T., Kiupel, M., Maier, A & Klopfleisch, R. (2020). Computerized calculation of mitotic count distribution in canine cutaneous mast cell tumor sections: mitotic count is area dependent. *Veterinary Pathology, 57*(2), 214-226. doi:Artn 0300985819890686

Bertram, C. A., & Klopfleisch, R. (2017). The Pathologist 2.0: An update on digital pathology in veterinary medicine. *Vet Pathol, 54*(5), 756-766. doi:10.1177/0300985817709888

Brochhausen, C., Winther, H. B., Hundt, C., Schmitt, V. H., Schomer, E., & Kirkpatrick, C. J. (2015). A virtual microscope for academic medical education: the pate project. *Interact J Med Res, 4*(2), e11. doi:10.2196/ijmr.3495

Bueno-de-Mesquita, J. M., Nuyten, D. S. A., Wesseling, J., van Tinteren, H., Linn, S. C., & van de Vijver, M. J. (2010). The impact of inter-observer variation in pathological assessment of node-negative breast cancer on clinical risk assessment and patient selection for adjuvant systemic treatment. *Annals of Oncology, 21*(1), 40-47. doi:10.1093/annonc/mdp273

Bulten, W., Hulsbergen-van de Kaa, C. A., van der Laak, J., & Litjens, G. J. (2018). *Automated segmentation of epithelial tissue in prostatectomy slides using deep learning.* Paper presented at the Medical Imaging 2018: Digital Pathology.

Cangul, T. (2001). Improved classification, diagnosis and prognosis of canine round cell tumours. *Veterinary Sciences Tomorrow*. doi:Urn:nbn:nl:ui:10-1874-7150

Cerezo-Echevarria, A., Grassinger, J. M., Beitzinger, C., Klopfleisch, R., & Aupperle-Lellbach, H. (2020). Evaluating the histologic grade of digital squamous cell carcinomas in dogs with dark and light haircoat-a comparative study of the invasive front and tumor cell budding systems. *Vet Sci, 8*(1). doi:10.3390/vetsci8010003

Chan, H. P., Hadjiiski, L. M., & Samala, R. K. (2020). Computer-aided diagnosis in the era of deep learning. *Medical Physics, 47*(5), e218-e227.

Chandradevan, R., Aljudi, A. A., Drumheller, B. R., Kunananthaseelan, N., Amgad, M., Gutman, D. A., Cooper, L. A. D. & Jaye, D. L. (2020). Machine-based detection and classification for bone marrow aspirate differential counts: initial development focusing on nonneoplastic cells. *Laboratory Investigation,100*(1), 98-109. doi:10.1038/s41374-019-0325-7

Chen, C.-L., Chen, C.-C., Yu, W.-H., Chen, S.-H., Chang, Y.-C., Hsu, T.-I., Hsiao, M., Yeh, C-Y & Chen, C.

(2021). An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. *Nature Communications, 12*.

Chen, G., Liu, H., Yu, L., Wei, Q., & Zhang, X. (2006). A new approach to classification based on association rule mining. *Decision Support Systems, 42*(2), 674-689.

Collins, J. T., Knapper, J., Stirling, J., Mduda, J., Mkindi, C., Mayagaya, V., race A. Mwakajinga, Nyakyi, P.T. Sanga, V. L., Carbery, D., White, L., Dale, S., Lim, Z., Baumberg, J. J., Cicuta, P., McDermott, S., Vodenicharski, B. & Bowman, R. (2020). Robotic microscopy for everyone: the OpenFlexure microscope. *Biomedical Optics Express, 11*(5), 2447-2460. doi:10.1364/BOE.385729

Confalonieri, R., Coba, L., Wagner, B., & Besold, T. R. (2021). A historical perspective of explainable Artificial Intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11*(1), e1391.

Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A. L., Razavian, N. & Tsirigos, A. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology imagesusing deep learning. *Nature medicine, 24*(10), 1559-1567.

Dario, P., Guglielmelli, E., & Allotta, B. (1994). *Robotics in medicine.* Paper presented at the Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'94).

De Chaumont, F., Dallongeville, S., Chenouard, N., Hervé, N., Pop, S., Provoost, T., Le Montagner, Y. Pankajakshan, P., Lecomte, T., Montagner, Y. L., Lagache, T., Dufour, A. & Olivo-Marin, J-C. (2012). Icy: an open bioimage informatics platform for extended reproducible research. *Nature methods, 9*(7), 690-696.

De Vera Mudry, M. C., Martin, J., Schumacher, V., & Venugopal, R. (2021). Deep learning in toxicologic pathology: a new approach to evaluate rodent retinal atrophy. *Toxicologic Pathology, 49*(4), 851-861. doi:10.1177/0192623320980674

de Wit, N. J., van Muijen, G. N., & Ruiter, D. J. (2004). Immunohistochemistry in melanocytic proliferative lesions. *Histopathology, 44*(6), 517-541. doi:10.1111/j.1365-2559.2004.01860.x

Dee, F. R. (2009). Virtual microscopy in pathology education. *Hum Pathol, 40*(8), 1112-1121. doi:10.1016/j.humpath.2009.04.010

Dee, F. R., & Fales-Williams, A. (2005). Virtual slidebox of comparative cancer pathology. *Faseb Journal, 19*(4), A244-A244.

Dennis, M. M., McSporran, K. D., Bacon, N. J., Schulman, F. Y., Foster, R. A., & Powers, B. E. (2011). Prognostic factors for cutaneous and subcutaneous soft tissue sarcomas in dogs. *Vet Pathol, 48*(1), 73-84. doi:10.1177/0300985810388820

Dimitriou, N., Arandjelovic, O., & Caie, P. D. (2019). Deep Learning for Whole Slide Image Analysis: An Overview. *Front Med (Lausanne), 6*, 264. doi:10.3389/fmed.2019.00264

Doi, K. (2007). Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics, 31*(4-5), 198-211. doi:10.1016/j.compmedimag.2007.02.002

Donalek, C. (2011). *Supervised and unsupervised learning.* Paper presented at the Astronomy Colloquia. USA.

Dorn, C. R. (1967). The epidemiology of cancer in animals. *Calif Med, 107*(6), 481-489.

Dorn, C. R., Taylor, D. O., Schneider, R., Hibbard, H. H., & Klauber, M. R. (1968). Survey of animal neoplasms in Alameda and Contra Costa Counties, California. II. Cancer morbidity in dogs and cats from Alameda County. *J Natl Cancer Inst, 40*(2), 307-318.

Dunn, B. E., Choi, H., Recla, D. L., Kerr, S. E., & Wagenman, B. L. (2009). Robotic surgical telepathology between the Iron Mountain and Milwaukee Department of Veterans Affairs Medical Centers: a twelve year experience. *Semin Diagn Pathol, 26*(4), 187-193. doi:10.1053/j.semdp.2009.09.007

Ehteshami Bejnordi, B., Mullooly, M., Pfeiffer, R. M., Fan, S., Vacek, P. M., Weaver, D. L., Karssemeijer, N., Beck, A. H., Gierach, G. L., van der Laak, J. A. W. M. & Sherman, M. E. (2018). Using deep convolutional neural networks to identify and classify tumor-associated stroma in diagnostic breast biopsies. *Modern Pathology, 31*(10), 1502-1512.

Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J., the CAMELYON16 Consortium, Hermsen, M., Manson, Q. F., Balkenhol, M., Geessink, O., Stathonikos, N., van Dijk, M. C., Bult, P., Beca, F., Beck, A. H., Wang, D., Khosla, A., Gargeya, R., Venancio, R. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA, 318*(22), 2199-2210. doi:10.1001/jama.2017.14585

Ertosun, M. G., & Rubin, D. L. (2015). *Automated grading of gliomas using deep learning in digital pathology images: a modular approach with ensemble of convolutional neural networks.* Paper presented at the AMIA annual symposium proceedings.

Evans, A. J., Chetty, R., Clarke, B. A., Croul, S., Ghazarian, D. M., Kiehl, T. R., T. R., Ordonez, B. P., Ilaalagan, S., & Asa, S. L. (2009). Primaryfrozen section diagnosis by robotic microscopy and virtual slide telepathology: the University Health Network experience. *Semin Diagn Pathol, 26*(4), 165-176. doi:10.1053/j.semdp.2009.09.006

Farahani, N., & Pantanowitz, L. (2016). Overview of Telepathology. *Clinics in Laboratory Medicine, 36*(1), 101-112. doi:10.1016/j.cll.2015.09.010

Farahani, N., Riben, M., Evans, A. J., & Pantanowitz, L. (2016). International Telepathology: Promises and Pitfalls. *Pathobiology, 83*(2-3), 121-126. doi:10.1159/000442390

Ferreira, R., Moon, B., Humphries, J., Sussman, A., Saltz, J., Miller, R., & Demarzo, A. (1997). The Virtual Microscope. *Proc AMIA Annu Fall Symp*, 449-453.

Foersch, S., Eckstein, M., Wagner, D. C., Gach, F., Woerl, A. C., Geiger, J., Glasner, C., Schelbert, S., Schulz, S., Porubsky, S., Kreft, A., Hartmann, A., Agaimy, A., & Roth, W. (2021). Deep learning for diagnosis and survival prediction in soft tissue sarcoma. *Annals of Oncology, 32*(9),1178-1187. doi:10.1016/j.annonc.2021.06.007

Fu, Y., Xue, P., Ji, H. Z., Cui, W. T., & Dong, E. Q. (2020). Deep model with Siamese network for viable and necrotic tumor regions assessment in osteosarcoma. *Medical Physics, 47*(10), 4895-4905. doi:10.1002/mp.14397

Fujita, H. (2020). AI-based computer-aided diagnosis (AI-CAD): the latest review to read first. *Radiological physics and technology, 13*(1), 6-19.

Fung, G. (2001). A comprehensive overview of basic clustering algorithms.

Gamlem, H., Nordstoga, K., & Glattre, E. (2008). Canine neoplasia--introductory paper. *APMIS Suppl*(125), 5-18. doi:10.1111/j.1600-0463.2008.125m2.x

Garnelo, M., & Shanahan, M. (2019). Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. *Current Opinion in Behavioral Sciences, 29*, 17-23. doi:https://doi.org/10.1016/j.cobeha.2018.12.010

Gertych, A., Swiderska-Chadaj, Z., Ma, Z., Markiewicz, T., Cierniak, S., Salemi, H., Guzman, S., Walts, A. E., & Knudsen, B. S. (2019). Convolutional neural networks can accurately distinguish four histologic growth patternsof lung adenocarcinoma in digital slides. *Scientific Reports, 9*(1), 1-12.

Giger, M. L., & Suzuki, K. (2008). Computer-aided diagnosis. In *Biomedical information technology* (pp. 359-XXII): Elsevier.

Goldberg, I. G., Allan, C., Burel, J.-M., Creager, D., Falconi, A., Hochheiser, H., Johnston, J., Mellen, J., Sorger, P. K., & Swedlow, J. R. (2005). The Open Microscopy Environment (OME) Data Model and XML file: open tools for informaticsand quantitative analysis in biological imaging. *Genome biology, 6*(5), 1-13.

Goldschmidt, M. H. (1985). Benign and malignant melanocytic neoplasms of domestic animals. *Am J Dermatopathol, 7 Suppl*, 203-212. doi:10.1097/00000372-198501001-00039

Goldschmidt, M. H., Kiupel, M., Klopfleisch, R., Munday, J. S., & Sruggs, J. L. (2018). S*urgical pathology of tumors of domestic animals volume 1: volume 1: epithelial tumors of the skin: epithelial tumorsof the skin*: Davis Thompson Foundation.

Goldschmidt, M. H., & Pathology, A. R. o. (1998). *Histological classification of epithelial and melanocytic tumors of the skin of domestic animals*: Armed Forces Institute of Pathology.

Graf, R., Pospischil, A., Guscetti, F., Meier, D., Welle, M., & Dettwiler, M. (2018). Cutaneous tumors in swiss dogs: retrospective data from the swiss canine cancer registry, 2008-2013. *Vet Pathol, 55*(6), 809-820. doi:10.1177/0300985818789466

Graham, A. R., Bhattacharyya, A. K., Scott, K. M., Lian, F., Grasso, L. L., Richter, L. C., Carpenter, J. B., Chiang, S., Henderson, J. T., Lopez, A. M., Barker, G. P., & Weinstein, R. S. (2009). Virtual slide telepathology for an academic teaching hospital surgical pathology qualityassurance program. *Hum Pathol, 40*(8), 1129-1136. doi:10.1016/j.humpath.2009.04.008

Grefenstette, J. J. (1993). *Genetic algorithms and machine learning.* Paper presented at the Proceedings of the sixth annual conference on Computational learning theory.

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell, 144*(5), 646-674. doi:10.1016/j.cell.2011.02.013

Harris, T. J. R., & McCormick, F. (2010). The molecular pathology of cancer. *Nature Reviews Clinical Oncology, 7*(5), 251-265. doi:10.1038/nrclinonc.2010.41

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778.

Heinemann, F., Birk, G., Schoenberger, T., & Stierstorfer, B. (2018). Deep neural network based histological scoring of lung fibrosis and inflammation in the mouse model system. *Plos One, 13*(8), e0202708. doi:10.1371/journal.pone.0202708

Hendrick, M. J. (1998). *Histological classification of mesenchymal tumors of skin and soft tissues of domestic animals*: Armed Forces Institute of Pathology.

Ho, D. J., Agaram, N. P., Schueffler, P. J., Vanderbilt, C. M., Jean, M.-H., Hameed, M. R., & Fuchs, T. J. (2020). *Deep interactive learning: an efficient labeling approach for deep learning-based osteosarcoma treatment response assessment.* Paper presented at the MICCAI.

Horvitz, E. J., Breese, J. S., & Henrion, M. (1988). Decision theory in expert systems and artificial intelligence. *International journal of approximate reasoning, 2*(3), 247-302.

Irshad, H., Montaser-Kouhsari, L., Waltz, G., Bucur, O., Nowak, J. A., Dong, F., Knoblauch, N. W., & Beck, A. H. (2015). Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: evaluating experts, automated methods, and the crowd. *Pac Symp Biocomput*, 294-305. doi:10.1142/9789814644730_0029

Jahn, S. W., Plass, M., & Moinfar, F. (2020). Digital Pathology: advantages, limitations and emerging perspectives. *J Clin Med, 9*(11). doi:10.3390/jcm9113697

Jiang, Y., Chen, L., Zhang, H., & Xiao, X. (2019). Breast cancer histopathological image classification using convolutional neural networks with small SE-ResNet module. *Plos One, 14*(3), e0214587.

Jiang, Y., Yang, M., Wang, S., Li, X., & Sun, Y. (2020). Emerging role of deep learning-based artificial intelligence in tumor pathology. *Cancer communications, 40*(4), 154-166.

Jones-Hall, Y. L., Skelton, J. M., & Adams, L. G. Implementing digital pathology into veterinary academics and research. *Journal of Veterinary Medical Education, 0*(0), e20210068. doi:10.3138/jvme-2021-0068

Jungbluth, A. A. (2008). Serological reagents for the immunohistochemical analysis of melanoma metastases in sentinel lymph nodes. *Semin Diagn Pathol, 25*(2), 120-125. doi:10.1053/j.semdp.2008.05.002

Kainz, P., Pfeiffer, M., & Urschler, M. (2017). Segmentation and classification of colon glands with deep convolutional neural networks and total variation regularization. *PeerJ, 5*, e3874.

Kang, J. W., Shin, S. H., Choi, J. H., Moon, K. C., Koh, J. S., Jung, C. K., Park, Y. K., Lee, K, B. & Chung, Y. G. (2017). Inter-and intra-observer reliability in histologic evaluation of necrosis rate induced by neo-adjuvant chemotherapy for osteosarcoma. *International Journal of Clinical and Experimental Pathology,10*(1), 359-367.

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons, 62*(1), 15-25. doi:10.1016/j.bushor.2018.08.004

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2021). Trust in Artificial Intelligence: Meta-analytic findings. *Hum Factors*, 187208211013988. doi:10.1177/00187208211013988

Kaplan, K. J., Burgess, J. R., Sandberg, G. D., Myers, C. P., Bigott, T. R., & Greenspan, R. B. (2002). Use of robotic Telepathology for frozen-section diagnosis: A retrospective trial of a telepathology system for intraoperative consultation. *Modern Pathology, 15*(11), 1197-1204. doi:10.1097/01.Mp.0000033928.11585.42

Keskinbora, K. H. (2019). Medical ethics considerations on artificial intelligence. *Journal of clinical neuroscience, 64*, 277-282.

Khameneh, F. D., Razavi, S., & Kamasak, M. (2019). Automated segmentation of cell membranes to evaluate HER2 status in whole slide images using a modified deep learning network. *Computers in Biology and Medicine, 110*, 164-174.

Kim, D., Pantanowitz, L., Schüffler, P., Yarlagadda, D. V. K., Ardon, O., Reuter, V. E., Hameed, M., Klimstra, D. S., & Hanna, M. G. (2020). (Re) Defining the high-power field for digital pathology. *Journal of Pathology Informatics,11*.

Kishore, D. R., & Kaur, T. (2012). *Backpropagation Algorithm: An Artificial Neural Network approach for pattern recognition*.

Koenig, A., Wojcieszyn, J., Weeks, B. R., & Modiano, J. F. (2001). Expression of S100a, vimentin, NSE, and melan A/MART-1 in seven canine melanoma cells lines and twenty-nine retrospective cases of canine melanoma. *Vet Pathol, 38*(4), 427-435. doi:10.1354/vp.38-4-427

Kok, M. K., Chambers, J. K., Tsuboi, M., Nishimura, R., Tsujimoto, H., Uchida, K., & Nakayama, H. (2019). Retrospective study of canine cutaneous tumors in Japan, 2008-2017. *J Vet Med Sci, 81*(8), 1133-1143. doi:10.1292/jvms.19-0248

Kolles, H., Wangenheim, A. v., Vince, G. H., Niedermayer, I., & Feiden, W. (1995). Automated grading of astrocytomas based on histomorphometric analysis of Ki-67 and Feulgen stained paraffin sections. Classification results of neuronal networks and discriminant analysis. *Analytical cellular pathology, 8*(2), 101-116.

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine, 23*(1), 89-109.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering, 160*(1), 3-24.

Krogh, A. (2008). What are artificial neural networks? *Nature biotechnology, 26*(2), 195-197.

Kulkarni, A., Chong, D., & Batarseh, F. A. (2020). Foundations of data imbalance and solutions for a data democracy. In *data democracy* (pp. 83-106): Elsevier.

Kumar, A., Singh, S. K., Saxena, S., Lakshmanan, K., Sangaiah, A. K., Chauhan, H., Shrivastava, S. & Singh, R. K. (2020). Deep feature learning for histopathological image classification of canine mammary tumors andhuman breast cancer. *Information Sciences, 508*, 405-421. doi:https://doi.org/10.1016/j.ins.2019.08.072

Kumar, R. K., Velan, G. M., Korell, S. O., Kandara, M., Dee, F. R., & Wakefield, D. (2004). Virtual microscopy for learning and assessment in pathology. *Journal of Pathology, 204*(5), 613-618. doi:10.1002/path.1658

Kuntz, C. A., Dernell, W. S., Powers, B. E., Devitt, C., Straw, R. C., & Withrow, S. J. (1997). Prognostic factors for surgical treatment of soft-tissue sarcomas in dogs: 75 cases (1986-1996). *J Am Vet Med Assoc, 211*(9), 1147-1151.

Lamprecht, M. R., Sabatini, D. M., & Carpenter, A. E. (2007). CellProfiler™: free, versatile software for automated biological image analysis. *Biotechniques, 42*(1), 71-75.

Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., Armañanzas, R., Santafé, G., Pérez, A., & Robles, V. (2006). Machine learningin bioinformatics. *Briefings in bioinformatics, 7*(1), 86-112.

Lee, J.-G., Jun, S., Cho, Y.-W., Lee, H., Kim, G. B., Seo, J. B., & Kim, N. (2017). Deep learning in medical imaging: general overview. *Korean journal of radiology, 18*(4), 570-584.

Lee, J. H., Shin, J., & Realff, M. J. (2018). Machine learning: Overview of the recent progresses and implications for the process systems engineering field. *Computers & Chemical Engineering, 114*, 111-121.

Leonard, J., & Kramer, M. A. (1990). Improvement of the backpropagation algorithm for training neural networks. *Computers & Chemical Engineering, 14*(3), 337-341. doi:https://doi.org/10.1016/0098-1354(90)87070-6

Li, J. X., Garfinkel, J., Zhang, X. R., Wu, D., Zhang, Y. J., de Haan, K., Wang, H., Liu, T., Bai, B., Rivenson, Y., Rubinstein, G., Scumpia, P. O., & Ozcan, A. (2021). Biopsy-free in vivo virtual histology of skin using deep learning. *Light-Science & Applications, 10*(1). doi:ARTN23310.1038/s41377-021-00674-8

Lim, C. P., Woo, S. C., Loh, A. S., & Osman, R. (2000). *Speech recognition using artificial neural networks.* Paper presented at the Proceedings of the First International Conference on Web Information Systems Engineering.

Lindman, K., Rose, J. F., Lindvall, M., Lundstrom, C., & Treanor, D. (2019). Annotations, ontologies, and Whole Slide Images - development of an annotated ontology-driven Whole Slide Image library of normal and abnormal human tissue. *J Pathol Inform, 10*, 22. doi:10.4103/jpi.jpi_81_18

Litjens, G., Sánchez, C. I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., Hulsbergen-van de Kaa, C., Bult, P., van Ginneken, B., & van der Laak, J. (2016). Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific Reports, 6*(1), 1-11.

Liu, J. X., Xu, B. L., Zheng, C., Gong, Y. H., Garibaldi, J., Soria, D., Green, A., Ellis, I. O., Zou, W., & Qiu, G. (2019). An end-to-end deeplearning histochemical scoring system for breast cancer TMA. *IEEE Transactions on Medical Imaging, 38*(2), 617-628. doi:10.1109/Tmi.2018.2868333

Liu, Q., Zhang, N., Yang, W., Wang, S., Cui, Z., Chen, X., & Chen, L. (2017). *A review of image recognition with deep convolutional neural network.* Paper presented at the International conference on intelligent computing.

Lopez-Rubio, E., Elizondo, D. A., Grootveld, M., Jerez, J. M., & Luque-Baena, R. M. (2015). Computational intelligence techniques in medicine. *Computational and Mathematical Methods in Medicine, 2015*. doi:Artn 196976 10.1155/2015/196976

Marée, R., Rollus, L., Stévens, B., Hoyoux, R., Louppe, G., Vandaele, R., Begon, J. M., Kainz, P., Geurts, P., & Wehenkel, L. (2016). Collaborative analysis of multi-gigapixel imaging data using Cytomine. *Bioinformatics, 32*(9),1395-1401.

Marino, D. J., Matthiesen, D. T., Stefanacci, J. D., & Moroff, S. D. (1995). Evaluation of dogs with digit masses: 117 cases (1981-1991). *J Am Vet Med Assoc, 207*(6), 726-728.

Marzahl, C., Aubreville, M., Bertram, C. A., Stayt, J., Jasensky, A. K., Bartenschlager, F., Fragoso-Garcia, M., Barton, A. K., Elsemann, S., Jabari, S., Krauth, J., Madhu, P., Voigt, J., Hill, J., Klopfleisch, R., & Maier, A. (2020). Deep Learning-based quantification of pulmonary hemosiderophages in cytology slides.*Scientific Reports, 10*(1). doi:ARTN 9795 10.1038/s41598-020-65958-2

Massone, C., Wurm, E. M., Hofmann-Wellenhof, R., & Soyer, H. P. (2008). Teledermatology: an update. *Semin Cutan Med Surg, 27*(1), 101-105. doi:10.1016/j.sder.2007.12.002

McCarthy, J. (2007). What is artificial intelligence?

McCullough, B., Ying, X., Monticello, T., & Bonnefoi, M. (2004). Digital microscopy imaging and new approaches in toxicologic pathology. *Toxicologic Pathology, 32 Suppl 2*, 49-58. doi:10.1080/01926230490451734

McHale, B., Blas-Machado, U., Oliveira, F. N., & Rissi, D. R. (2018). A divergent pseudoglandular configuration of cutaneous plasmacytoma in dogs. *Journal of Veterinary Diagnostic Investigation, 30*(2), 260-262. doi:10.1177/1040638717735868

Merlo, D. F., Rossi, L., Pellegrino, C., Ceppi, M., Cardellino, U., Capurro, C., Ratto, A., Sambucco, P. L., Sestito, V., Tanara, G., & Bocchini, V. (2008). Cancerincidence in pet dogs: findings of the Animal Tumor Registry of Genoa, Italy. *J Vet Intern Med, 22*(4), 976-984. doi:10.1111/j.1939-1676.2008.0133.x

Meuten, D. J. (2016). *Tumors in Domestic Animals*: Wiley.

Meyer, D., & Wien, F. (2015). Support vector machines. *The Interface to libsvm in package e1071, 28*.

Nam, S., Chong, Y., Jung, C. K., Kwak, T. Y., Lee, J. Y., Park, J., Rho, M. J., & Go, H. (2020). Introduction to digitalpathology and computer-aided pathology. *J Pathol Transl Med, 54*(2), 125-134. doi:10.4132/jptm.2019.12.31

Narayanan, P. L., Raza, S. E. A., Dodson, A., Gusterson, B., Dowsett, M., & Yuan, Y. (2018). DeepSDCS: Dissecting cancer proliferation heterogeneity in Ki67 digital whole slide images. *arXiv preprint arXiv:1806.10850*.

Navada, A., Ansari, A. N., Patil, S., & Sonkamble, B. A. (2011). *Overview of use of decision tree algorithms in machine learning.* Paper presented at the 2011 IEEE control and system graduate research colloquium.

Nichols, J. A., Herbert Chan, H. W., & Baker, M. A. (2019). Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophysical reviews, 11*(1), 111-118.

Nordrum, I., Engum, B., Rinde, E., Finseth, A., Ericsson, H., Kearney, M., Stalsberg, H., & Eide, T. J. (1991). Remote frozen section service: a telepathology project in northern Norway. *Hum Pathol, 22*(6), 514-518.doi:10.1016/0046-8177(91)90226-f

Norgan, A., Hart, S. N., Flotte, W., Shah, K., Buchan, Z. R., Mounajjed, T., & Flotte, T. (2018). Classification of melanocytic lesions in selected and Whole Slide Images Using a Convolutional Neural Network. *Laboratory Investigation, 98*, 595-595.

Ohsie, S. J., Sarantopoulos, G. P., Cochran, A. J., & Binder, S. W. (2008). Immunohistochemical characteristics of melanoma. *J Cutan Pathol, 35*(5), 433-444. doi:10.1111/j.1600-0560.2007.00891.x

Orlando, L., Viale, G., Bria, E., Lutrino, E. S., Sperduti, I., Carbognin, L., Schiavone, P., Quaranta, A., Fedele, P., Caliolo, C., Calvani, N., Criscuolo, M., & Cinieri, S. (2016). Discordancein pathology report after central pathology review: Implications for breast cancer adjuvant treatment. *Breast, 30*, 151-155. doi:10.1016/j.breast.2016.09.015

Pantanowitz, L., Sharma, A., Carter, A. B., Kurc, T., Sussman, A., & Saltz, J. (2018). Twenty years of digital pathology: an overview of the road travelled, what is on the horizon, and the emergence of vendor-neutral archives. *J Pathol Inform, 9*, 40. doi:10.4103/jpi.jpi_69_18

Patterson, D. (1990). *Introduction to artificial intelligence and expert systems*: Prentice-Hall, Inc.

Poole, D. L., & Mackworth, A. K. (2010). *Artificial Intelligence: foundations of computational agents*: Cambridge University Press.

Poynter, F. N. L. (1967). Marcello Malpighi and the evolution of embryology. *Medical History, 11*, 426 - 427.

Provost, F., & Kohavi, R. (1998). Glossary of terms. *Journal of Machine Learning, 30*(2-3), 271-274.

Ramos-Vara, J. A., & Miller, M. A. (2011). Immunohistochemical identification of canine melanocytic neoplasms with antibodies to melanocytic antigen PNL2 and tyrosinase: comparison with Melan A. *Veterinary Pathology, 48*(2), 443-450. doi:10.1177/0300985810382095

Ramos-Vara, J. A., Miller, M. A., & Valli, V. E. O. (2007). Immunohistochemical detection of multiple myeloma 1/interferon regulatory factor 4 (MUM1/IRF-4) in canine plasmacytoma: Comparison with CD79a and CD20. *Veterinary Pathology, 44*(6), 875-884. doi:DOI 10.1354/vp.44-6-875

Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-Net: Convolutional Networks for Biomedical Image*

Segmentation, Cham.

Roux, L., Racoceanu, D., Lomenie, N., Kulikova, M., Irshad, H., Klossa, J., Capron, F., Genestie, C., Le Naour, G., & Gurcan, M. N. (2013). Mitosisdetection in breast cancer histological images An ICPR 2012 contest. *J Pathol Inform, 4*, 8. doi:10.4103/2153-3539.112693

Rueden, C. T., Schindelin, J., Hiner, M. C., DeZonia, B. E., Walter, A. E., Arena, E. T., & Eliceiri, K. W. (2017). ImageJ2: ImageJ for the next generation of scientific image data. *Bmc Bioinformatics, 18*(1), 1-26.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision, 115*(3), 211-252.doi:10.1007/s11263-015-0816-y

Russell, S., & Norvig, P. (2002). Artificial intelligence: a modern approach.

Salvi, M., Molinari, F., Iussich, S., Muscatello, L. V., Pazzini, L., Benali, S., Banco, B., Abramo, F., De Maria, R., & Aresu, L. (2021). Histopathological classification of canine cutaneous round cell tumors using deep learning: amulti-center study. *Frontiers in Veterinary Science, 8*. doi:ARTN 640944 10.3389/fvets.2021.640944

Sandusky, G. E., Carlton, W. W., & Wightman, K. A. (1987). Diagnostic immunohistochemistry of canine round cell tumors. *Vet Pathol, 24*(6), 495-499. doi:10.1177/030098588702400604

Santana, C. H., Moreira, P. R. R., Rosolem, M. C., & Vasconcelos, R. O. (2016). Relationship between the inflammatory infiltrate and the degree of differentiation of the canine cutaneous squamous cell carcinoma. *Vet Anim Sci, 1-2*, 4-8. doi:10.1016/j.vas.2016.10.001

Schaumberg, A. J., Rubin, M. A., & Fuchs, T. J. (2017). H&E-stained whole slide image deep learning predicts SPOP mutation state in prostate cancer. *BioRxiv*, 064279.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J. Y., White, D. J., Hartenstein, V., Eliceiri, K., Tomancak, P., & Cardona, A. (2012). Fiji: an open-source platform for biological-image analysis. *Nature methods, 9*(7), 676-682.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks, 61*, 85-117.

Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature methods, 9*(7), 671-675.

Schnitt, S. J. (2001). Traditional and newer pathologic factors. *J Natl Cancer Inst Monogr*(30), 22-26. doi:10.1093/oxfordjournals.jncimonographs.a003456

Sheehan, S. M., & Korstanje, R. (2018). Automatic glomerular identification and quantification of histological phenotypes using image analysis and machine learning. *Am J Physiol Renal Physiol, 315*(6), F1644-f1651. doi:10.1152/ajprenal.00629.2017

Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. *Ieee Access, 7*, 53040-53065. doi:10.1109/ACCESS.2019.2912200

Siddique, N., & Adeli, H. (2013). *Computational intelligence: synergies of fuzzy logic, neural networks and evolutionary computing*: John Wiley & Sons.

Sidey-Gibbons, J. A., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction. *BMC medical research methodology, 19*(1), 1-18.

Siegel, G., Regelman, D., Maronpot, R., Rosenstock, M., Hayashi, S. M., & Nyska, A. (2018). Utilizing novel telepathology system in preclinical studies and peer review. *J Toxicol Pathol, 31*(4), 315-319. doi:10.1293/tox.2018-0032

Smedley, R. C., Lamoureux, J., Sledge, D. G., & Kiupel, M. (2011). Immunohistochemical diagnosis of canine oral amelanotic melanocytic neoplasms. *Veterinary Pathology, 48*(1), 32-40. doi:10.1177/0300985810387447

Smith, S. H., Goldschmidt, M. H., & McManus, P. M. (2002). A comparative review of melanocytic neoplasms. *Vet Pathol, 39*(6), 651-678. doi:10.1354/vp.39-6-651

Sommer, C., Straehle, C., Koethe, U., & Hamprecht, F. A. (2011). *Ilastik: Interactive learning and segmentation toolkit.* Paper presented at the 2011 IEEE international symposium on biomedical imaging: From nano to macro.

Steiner, D. F., MacDonald, R., Liu, Y., Truszkowski, P., Hipp, J. D., Gammage, C., Thng, F., Peng, L., & Stumpe, M. C. (2018).Impact of deep learning assistance on the histopathologic review of lymph nodes for metastaticbreast cancer. *The American journal of surgical pathology, 42*(12), 1636.

Stöger, K., Schneeberger, D., & Holzinger, A. (2021). Medical artificial intelligence: the European legal perspective. *Communications of the ACM, 64*(11), 34-36.

Sudharshan, P. J., Petitjean, C., Spanhol, F., Oliveira, L. E., Heutte, L., & Honeine, P. (2019). Multiple instance learning for histopathological breast cancer image classification. *Expert Systems with Applications, 117*, 103-111. doi:10.1016/j.eswa.2018.09.049

Sutton, R. S. (1992). Introduction: The challenge of reinforcement learning. In *Reinforcement Learning* (pp. 1-3): Springer.

Suzuki, K. (2017). Overview of deep learning in medical imaging. *Radiological physics and technology, 10*(3), 257-273. doi:10.1007/s12194-017-0406-5

Takeyas, B. L. (2007). Introducción a la inteligencia artificial. *Instituto Tecnológico de Nuevo Laredo. Web del autor: http://www.itnuevolaredo. edu. mx/takeyas*.

Tan, M., & Le, Q. V. (2019). EfficientNet: rethinking model scaling for convolutional neural networks. *ArXiv, abs/1905.11946*.

Teramoto, A., Tsukamoto, T., Kiriyama, Y., & Fujita, H. (2017). Automated classification of lung cancer types from cytological images using deep convolutional neural networks. *BioMed research international, 2017*.

Thrall, M. J., Wimmer, J. L., & Schwartz, M. R. (2015). Validation of multiple whole slide imaging scanners based on the guideline from the College of American Pathologists Pathology and Laboratory Quality Center. *Archives of Pathology & Laboratory Medicine, 139*(5), 656-664. doi:10.5858/arpa.2014-0073-OA

Ting, K. M. (2010). *Confusion Matrix.* Paper presented at the Encyclopedia of Machine Learning.

Tolkach, Y., Dohmgörgen, T., Toma, M., & Kristiansen, G. (2020). High-accuracy prostate cancer pathology using deep learning. *Nature Machine Intelligence, 2*(7), 411-418.

Turk, J. L. (1993). Virchow,Rudolf - father of cellular pathology. *Journal of the Royal Society of Medicine, 86*(12), 688-689.

Turner, O. C., Aeffner, F., Bangari, D. S., High, W., Knight, B., Forest, T., Cossic, B., Himmel, L. E., Rudmann, D. G., Bawa, B., Muthuswamy, A., Aina, O. H., Edmondson, E. F., Saravanan, C., Brown, D. L., Sing, T., & Sebastian, M. M. (2020). Society of toxicologic pathology digital pathology and image analysis special interest grouparticle*: opinion on the application of artificial intelligence and machine learning to digital toxicologic pathology. *Toxicologic Pathology, 48*(2), 277-294. doi:Artn 0192623319881401 10.1177/0192623319881401

Vandenberghe, M. E., Scott, M. L., Scorer, P. W., Söderberg, M., Balcerzak, D., & Barker, C. (2017). Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. *Scientific Reports, 7*(1), 1-11.

Veta, M., Van Diest, P. J., Willems, S. M., Wang, H., Madabhushi, A., Cruz-Roa, A., Gonzalez, F., Larsen, A. B., Vestergaard, J. S., Dahl, A. B., Cireşan, D. C., Schmidhuber, J., Giusti, A., Gambardella, L. M., Tek, F. B., Walter, T., Wang, C. W., Kondo, S., Matuszewski, B. J., Precioso, F., Pluim, J. P. (2015). Assessment of algorithms for mitosis detection in breast cancer histopathology images. *MedicalImage Analysis, 20*(1), 237-248.

Wang, D., Khosla, A., Gargeya, R., Irshad, H., & Beck, A. H. (2016). Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*.

Wang, L., Ding, L., Liu, Z., Sun, L., Chen, L., Jia, R., . . . Ye, J. (2020). Automated identification of malignancy in whole-slide pathological images: identification of eyelid malignant melanoma in gigapixel pathological slides using deep learning. *British Journal of Ophthalmology, 104*(3), 318-323. doi:10.1136/bjophthalmol-2018-313706

Wang, L. Y., Ding, L. Q., Liu, Z. F., Sun, L. L., Chen, L. R., Jia, R. B., Dai, X., Cao, J., & Ye, J. (2020). Automated identification of malignancy in whole-slide pathological images: identification of eyelid malignant melanoma in gigapixel pathological slides using deep learning. *British Journal of Ophthalmology, 104*(3), 318-323. doi:10.1136/bjophthalmol-2018-313706

Wang, S., Chen, A., Yang, L., Cai, L., Xie, Y., Fujimoto, J., Gazdar, A., & Xiao, G. (2018). Comprehensive analysis oflung cancer pathology images to discover tumor shape and boundary features that predict survival outcome. *Scientific Reports, 8*(1), 1-9.

Wang, S., Zhu, Y., Yu, L., Chen, H., Lin, H., Wan, X., Fan, X. & Heng, P.-A. (2019). RMDL: Recalibrated multi- instance deep learning for whole slide gastric image classification. *Medical Image Analysis, 58*,101549.

Weinstein, R. S. (1986). Prospects for telepathology. *Hum Pathol, 17*(5), 433-434. doi:10.1016/s0046-8177(86)80028-4

Weinstein, R. S., Bloom, K. J., & Rozek, L. S. (1987). Telepathology and the networking of pathology diagnostic services. *Archives of Pathology & Laboratory Medicine, 111*(7), 646-652.

Weinstein, R. S., Graham, A. R., Richter, L. C., Barker, G. P., Krupinski, E. A., Lopez, A. M., Erps, K. A., Bhattacharyya, A. K., Yagi, Y., & Gilbertson, J. R. (2009). Overview of telepathology, virtual microscopy, and whole slide imaging: prospects forthe future. *Hum Pathol, 40*(8), 1057-1069. doi:10.1016/j.humpath.2009.04.006

Wiener, D. J. (2021). Histologic features of hair follicle neoplasms and cysts in dogs and cats: a diagnostic guide. *Journal of Veterinary Diagnostic Investigation, 33*(3), 479-497. doi:Artn 1040638721993565

Wilm, F., Fragoso, M., Marzahl, C., Qiu, J., Bertram, C. A., Klopfleisch, R., Maier, A., Breininger, K., Aubreville, M. (2022). Pan-Tumor CAnine cuTaneous Cancer Histology (CATCH) Dataset. *arXiv preprint arXiv:2201.11446*.

Wu, M., & Chen, L. (2015). *Image recognition based on deep learning.* Paper presented at the 2015 Chinese Automation Congress (CAC).

Xin, M., & Wang, Y. (2019). Research on image classification model based on deep convolution neural network. *EURASIP Journal on Image and Video Processing, 2019*(1), 40. doi:10.1186/s13640-019-0417-8

Xing, X., Zhao, X., Wei, H., & Li, Y. (2021). Diagnostic accuracy of different computer-aided diagnostic systems for prostate cancer based on magnetic resonance imaging: A systematic review with diagnostic meta-analysis. *Medicine, 100*(3), e23817. doi:10.1097/md.0000000000023817

Ye, Y., Sun, W. W., Xu, R. X., Selmic, L. E., & Sun, M. Z. (2021). Intraoperative assessment of canine soft tissue sarcoma by deep learning enhanced optical coherence tomography. *Veterinary and Comparative Oncology, 19*(4), 624-631. doi:10.1111/vco.12747

Zainab, H., Sultana, A., & Shaimaa. (2019). Stromal desmoplasia as a possible prognostic indicator in different grades of oral squamous cell carcinoma. *J Oral Maxillofac Pathol, 23*(3), 338-343. doi:10.4103/jomfp.JOMFP_136_19

Zarella, M. D., Bowman, D., Aeffner, F., Farahani, N., Xthona, A., Absar, S. F., S. F., Parwani, A., Bui, M., & Hartman, D. J. (2019). Apractical guide to whole slide imaging a white paper from the digital pathology association. *Archives of Pathology & Laboratory Medicine, 143*(2), 222-234. doi:10.5858/arpa.2018-0343-RA

Zhou, Y.-T., & Chellappa, R. (2012). *Artificial neural networks for computer vision* (Vol. 5): Springer Science & Business Media.

Zou, J., Han, Y., & So, S.-S. (2008). Overview of artificial neural networks. *Artificial Neural Networks*, 14-22.

Zuraw, A., & Aeffner, F. (2022). Whole-slide imaging, tissue image analysis, and artificial intelligence in veterinary pathology: An updated introduction and review. *Veterinary Pathology, 59*(1), 6-25. doi:10.1177/03009858211040484

9. Acknowledgements

This thesis has been written at the Institute of Veterinary Pathology of the Free University of Berlin, for which my first acknowledgement and thankfulness is dedicated to this amazing university and to this institute.

I would like to express my special gratitude to Prof. Dr. Robert Klopfleisch for the insightful supervision of this work, for his patience, for his kind confidence and for the numerous professional and not so professional talks.

I would also like to thank my colleagues at the Institute of Veterinary Pathology at FU-Berlin for making my time here more enjoyable and entertaining, as well as enriching with regard to learning in pathology, especially my colleagues and friends Christof Bertram, Florian Bartenschlager, Sophie Merz, Anne Petrick, Chloé Puglet and Paula Sapion.

I would also like to thank the working group of the pattern recognition laboratory of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), especially to Frauke Wilm for her impressive work in developing the algorithm, as well as her contagious passion for computer science in the medical sciences.

I would like to thank Ms. Nicole Huth for her competent technical support in the laboratory. I would like to thank Alexander Bartel of the Institute of Veterinary Epidemiology and Biometrics for his competent statistical advice.

A special thanks to my parents and my sisters who, despite being far away, always supported me and were my driving force during these years, I love you all!

I want to dedicate this work completely to my brother and best friend Oscar who was always with me in the distance and although now he is not physically in this world, his motivation and brotherly love is still latent within me. I love you brother! Thanks for everything!

Last but not least, to my faithful partner and great friend, my dog "Wolvie", who accompanied me since the beginning of this journey in my life and made it easier. I love you!

Founding sources

I would like to gratefully acknowledge the financial support of CONACyT (National Council for Science and Technology – Mexico) and the German Academic Exchange Service (DAAD) provided during the execution of this project.

Declaration of independence

I hereby confirm that I have written this thesis independently. I assure that I have used only the sources and aids indicated.

Conflicts of interest.

In the context of this work, there are no conflicts of interest due to donations from third parties.

Berlin, 25.11.2022 **Marco Antonio Fragoso García**