Aus der Medizinischen Klinik und Poliklinik III - Großhadern

Klinik der Ludwig-Maximilian-Universität München

Direktor: Prof. Dr. Dr. Michael von Bergwelt

An Automated Classification System for Leukocyte Morphology in Acute Myeloid Leukemia

Dissertation
zum Erwerb des Doktorgrades der Medizin
an der Medizinischen Fakultät der
Ludwig-Maximilians-Universität zu München

vorgelegt von
Christian Matek

aus
Amberg

2021

Mit Genehmigung der Medizinischen Fakultät der Universität München

Berichterstatter:	Prof. Dr. med. Karsten Spiekermann
Mitberichterstatter:	Prof. Dr. med. Tobias Feuchtinger
	Priv. Doz. Dr. rer. nat. Hanna-Mari Baldauf
	Prof. Dr. med. Dr. phil. Torsten Haferlach
Mitbetreuung durch den promovierten Mitarbeiter:	Dr. rer. nat. Carsten Marr
Dekan:	Prof. Dr. med. dent. Reinhard Hickel
Tag der mündlichen Prüfung:	30.09.2021

Eidesstattliche Versicherung

Ich erkläre hiermit an Eides statt,

dass ich die vorliegende Dissertation mit dem Titel

An Automated Classification System for Leukocyte Morphology in Acute Myeloid Leukemia

selbstständig verfasst, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

Ich erkläre des Weiteren, dass die hier vorgelegte Dissertation nicht in gleicher oder in ähnlicher Form bei einer anderen Stelle zur Erlangung eines akademischen Grades eingereicht wurde.

München, 01. Oktober 2021

Christian Matek

Christian Matek

Diese Arbeit beruht auf folgenden begutachteten Veröffentlichungen:

Publikationen

C. Matek, S. Schwarz, K. Spiekermann und C. Marr, Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks.
Nature Machine Intelligence **1**, 538–544 (2019)

C. Matek, S. Schwarz, K. Spiekermann und C. Marr, A single-cell image database of leukocyte morphologies relevant in Acute Myeloid Leukemia.
Scientific Data, under review (2020)

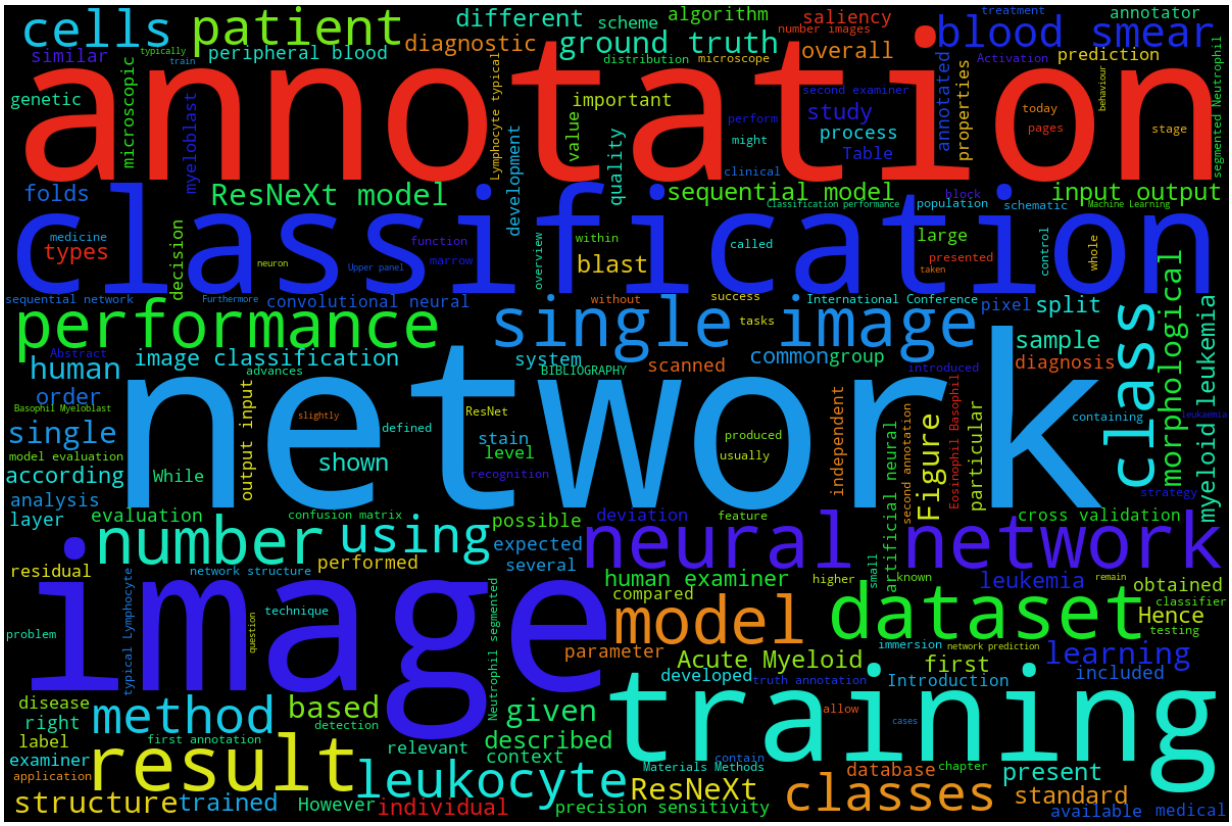
Datensatz

C. Matek, S. Schwarz, K. Spiekermann und C. Marr, A single-cell morphological dataset of leukocytes from AML patients and non-malignant controls (AML-Cytomorphology_LMU).
The Cancer Imaging Archive `https://doi.org/10.7937/tcia.2019.36f5o9ld` (2019)

Abstracts

Matek C., Marr C., Spiekermann K. (2018) Abstract: Digital Cytomorphology. In: Maier A., Deserno T., Handels H., Maier-Hein K., Palm C., Tolxdorff T. (eds) *Bildverarbeitung für die Medizin 2018. Informatik aktuell.* Springer Vieweg, Berlin, Heidelberg

Matek C., Schwarz S., Spiekermann K., Marr C. (2020) Recognition of AML Blast Cells in a Curated Single-Cell Dataset of Leukocyte Morphologies Using Deep Convolutional Neural Networks. In: Tolxdorff T., Deserno T., Handels H., Maier A., Maier-Hein K., Palm C. (eds) *Bildverarbeitung für die Medizin 2020. Informatik aktuell.* Springer Vieweg, Wiesbaden



Wordcloud, produced using word_cloud (https://amueller.github.io/word_cloud).

Contents

List of Figures

List of Tables

Zusammenfassung

In der Diagnostik hämatologischer Erkrankungen wie der akuten myeloischen Leukämie haben sich in den vergangenen Jahren bedeutende Fortschritte ergeben, die vor allem auf einem vertieften Verständnis ihrer biologischen und genetischen Ursachen beruhen. Trotzdem spielt die zytomorphologische Untersuchung von Blut- und Knochenmarkspräparaten nach wie vor eine zentrale Rolle in der diagnostischen Aufarbeitung. Die mikroskopische Begutachtung dieser Präparate konnte bisher nicht automatisiert werden und erfolgt nach wie vor durch menschliche Befunder, die eine manuelle Differentierung und Auszählung relevanter Zelltypen vornehmen. Daher ist der Zugang zu zytomorphologischen Untersuchungen durch die Zahl verfügbarer zytologischer Befunder begrenzt. Darüber hinaus beruht die Beurteilung der Präparate auf der individuellen Einschätzung der Befunder und ist somit von deren Ausbildung und Erfahrung abhängig, was eine standardisierte und quantitative Auswertung der Morphologie zusätzlich erschwert.

Ziel der vorliegenden Arbeit ist es, ein computerbasiertes System zu entwickeln, die die morphologische Differenzierung von Leukozyten unterstützt. Zu diesem Zweck wird auf in den letzten Jahren entwickelte leistungsfähige Algorithmen aus dem Bereich der Künstlichen Intelligenz, insbesondere des sogenannten Tiefen Lernens zurückgegriffen. In einem ersten Schritt des Projekts wurden periphere Blutausstriche von AML-Patienten und Kontrollen mit Methoden der digitalen Pathologie erfasst. Erfahrene Befunder aus dem Labor für Leukämiediagnostik am LMU-Klinikum München annotierten die digitalisierten Präparate und differenzierten sie in ein 15-klassiges, aus der Routinediagnostik stammendes Standardschema. Auf diese Weise wurde mit über 18,000 morphologisch annotierten Leukozyten der aktuell größte öffentlich verfügbare Datensatz relevanter Einzelzellbilder zusammengestellt.

In einer zweiten Phase des Projekts wurde dieser Datensatz verwendet, um Algorithmen vom Typ neuronaler Faltungsnetze zur Klassifikation von Einzelzellbilden zu trainieren. Eine Analyse ihrer Vorhersagen zeigt dass diese Netzwerke Einzelzellbilder der meisten Zellklassen sehr erfolgreich differenzieren können. Für falsch klassifizierte Bilder ähnelt ihr Fehlermuster dem menschlicher Befunder. Neben der Klassifikation einzelner Zellen erlauben die Netzwerke auch die Beantwortung gröberer, binärer Fragestellungen, etwa ob eine bestimmte Zelle blastären Charakter hat oder zu den morphologischen Klassen gehört die in einem peripheren Blutausstrich nicht unter physiologischen Bedingungen vorkommen. Bei diesen Fragen zeigen die Netzwerke eine ähnliche und leicht bessere Leistung als der menschliche Befunder. Die Ergebnisse dieser Arbeit illustrieren das Potential von Methoden der künstlichen Intelligenz auf dem Gebiet der Hämatologie und eröffnen Möglichkeiten zu ihrer Weiterentwicklung zu einem praktischen Hilfsmittel der Leukämiediagnostik.

Abstract

Diagnosis of hematological malignancies and of acute myeloid leukemia in particular have undergone wide-ranging advances in recent years, driven by an increasingly detailed knowledge of its underlying biological and genetic mechanisms. Nevertheless, cytomorphologic evaluation of samples of peripheral blood and bone marrow is still an integral part of the routine diagnostic workup. Microscopic analysis of these samples has so far defied automation and is still mainly performed by human cytologists manually classifying and counting relevant cell populations. Access to this diagnostic modality is therefore limited by the number and availability of educated cytologists. Furthermore, its results rest on judgments of examiners, which may vary according to their education and experience, rendering rigorous quantification and standardization of the method difficult.

In this thesis, an approach to cytomorphologic classification is presented that aims to harness recent advances in computational image classification for leukocyte differentiation using Deep Learning techniques that derive from the domain of Artificial Intelligence. In a first stage of the project, peripheral blood smear samples from both AML patients and controls were scanned using techniques from digital pathology. Experienced cytologists from the Laboratory of Leukemia Diagnostics at the LMU Klinikum annotated the digitized samples according to a scheme of 15 morphological categories derived from standard routine diagnostics. The resulting set of over 18,000 annotated single-cell images is the largest public database of leukocyte morphologies in leukemia available today.

In a second step, the compiled dataset was used to develop a neural network that is able to classify leukocytes into the standard morphological scheme. Evaluation of network predictions show that the network performs well at the classification task for most clinically relevant categories, with an error pattern similar to that of human examiners. The network can also be employed to answer two questions of immediate clinical relevance, namely if a given single-cell image shows a blast-like cell, or if it belongs to the set of atypical cells which are not present in peripheral blood smears under physiological conditions. At these questions, the network is found to show similar and slightly better performance compared to the human examiner. These results show the potential of Deep Learning techniques in the field of hematological diagnostics and suggest avenues for their further development as a helpful tool of leukemia diagnostics.

Abbreviations

AML	Acute Myeloid Leukaemia
AMMoL	Acute myelomonocytic Leukaemia
AOI	Area of Interest
APL	Acute Promyelocytic Leukaemia
ATO	Arsenic trioxide
ATRA	All-trans retinoic acid
AUC	Area under the Curve
CNN	Convolutional Neural Network
CNTK	Microsoft Cognitive Toolkit
DOF	Depth of Field
EDTA	Ethylenediamine tetraacetic acid
FAB	French-American-British
FISH	Fluorescence in-situ hybridization
FOV	Field of view
GPU	Graphics processing unit
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
MDS	Myelodysplastic Syndrome
MPN	Myeloproliferative Neoplasm
NA	Numerical Aperture
PAS	Periodic acid–Schiff
PCR	Polymerase Chain Reaction
ReLU	Rectified Linear Unit
ResNet	Residual Network

ROC	Receiver Operating Characteristic
TCIA	The Cancer Imaging Archive
VGG	Visual Geometry Group
WHO	World Health Organization

Chapter 1

Introduction

In this chapter, Acute Myeloid Leukemia (AML), the disease on which this work is focused, is introduced. Its definition, pathogenesis and clinical properties are briefly reviewed, and an overview of the most common diagnostic workup is presented, with an emphasis on microscopic examination of blood smears. A brief review of artificial neural networks is given, and their use in the context of image classification is explained. Finally, aims and scope of the project are defined.

1.1 Acute Myeloid Leukemia

1.1.1 History

The pathophysiologic relevance of peripheral leukocytosis in the absence of infection was first realized in 1845 by Bennett [1] and Virchow [2], who established the term leukemia (derived from the ancient greek terms for "white" and for "blood") for the high numbers of leukocytes he observed in peripheral blood. Later, Ebstein introduced the notion of acute leukemias to distinguish their rapidly progressive and quickly fatal clinical course from a more indolent, chronic form of the disease [3]. Finally, the distinction of myeloid and lymphoblastic leukemias was devised by Naegeli in 1900 [4], who also recognised the myeloblast as the key malignant cell type of myeloid leukemias. In summary, these findings allowed defining acute myeloid leukemia (AML) as a disease entity.

1.1.2 Etiology

In a present-day understanding, AML is a pathology based on the malignant transformation of cells belonging the hematopoietic system. Specifically, neoplastic cells involved in AML are part of the myeloid lineage of hematopoiesis (cf. Fig. 1.1). Various cell types on different stages of myelopoiesis can be affected, leading to considerable diversity in etiology, clinical presentation, and prognosis [5]. In this context, the predominant cell type affected is the hematopoietic stem cell.

Typically, malignant transformation is due to one or several genetic lesions in the hematopoietic system, which in adults is located in the bone marrow [5]. As a consequence, uncontrolled proliferation of immature myeloid progenitor cells takes place in the bone marrow, displacing other physiological cell types. In many cases, this uncontrolled proliferation leads to flooding of immature myeloid cells into the peripheral blood stream.

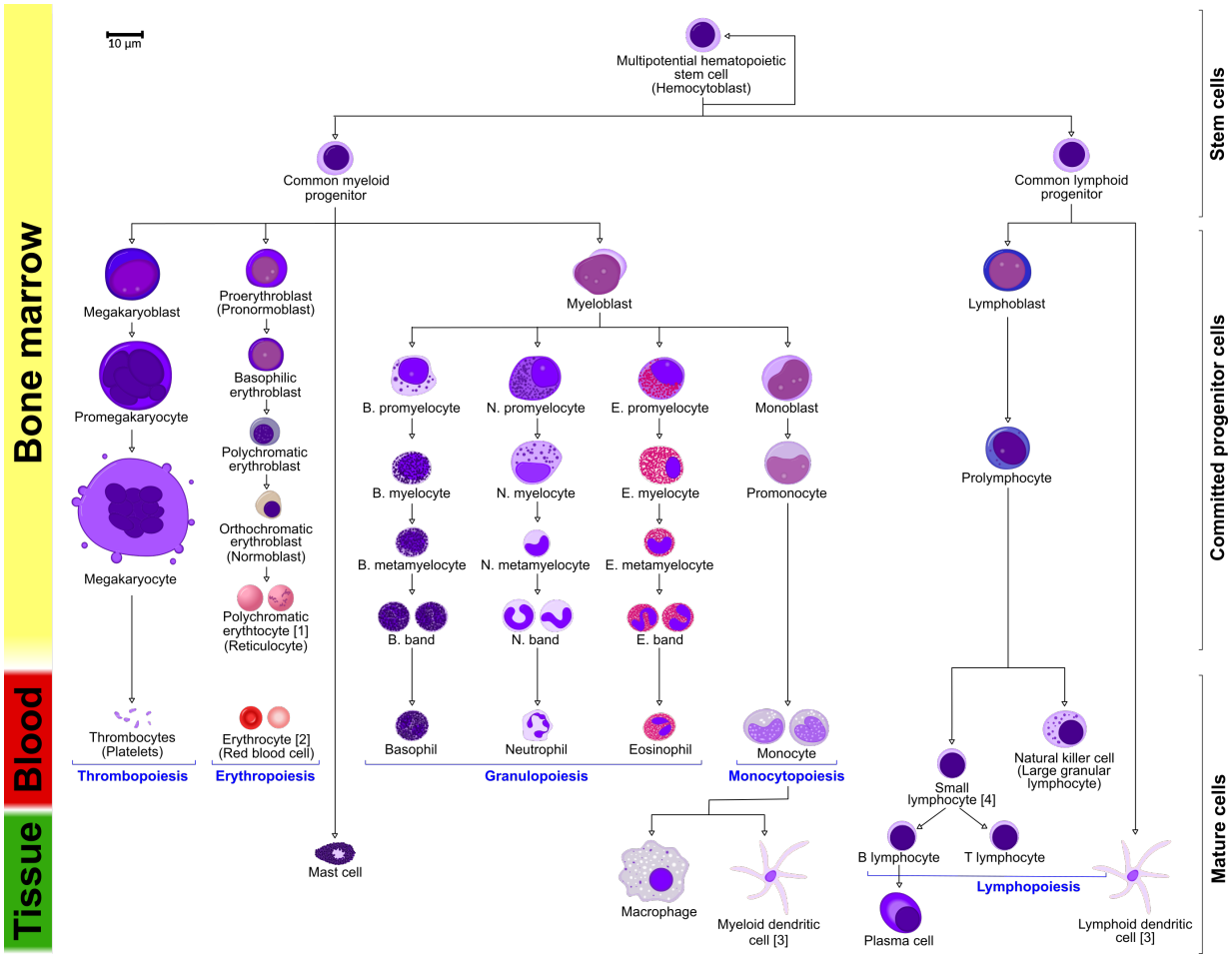


Figure 1.1: Simplified, schematic depiction of the morphological cell types and differentiation pathways involved in human hematopoiesis, including the myeloid and lymphoid lineage. Image reproduced from Ref. [6].

However, malignant proliferation of hematopoietic cells can also lead to a normal or reduced number of peripheral leukocytes [9], which is why their relative frequency in the bone marrow is relevant for definitive diagnosis. In this context, presence of immature blast cells above a threshold of 20% of nucleated cells in the bone marrow is required today to establish the diagnosis for most subtypes of AML [7].

Disease type	Defining properties
AML with recurrent genetic abnormalities	AML with t(8:21)(q22;q22); RUNX1-RUNX1T1 AML with inv(16)(p13.1q22) or t(16;16)(p13.1;q22); CBFB-MYH11 Acute promyelocytic leukemia (APL) with PML-RARA AML with t(9;11)(p21.3;q23.3); MLLT3-KMT2 AML with t(6;9)(p23;q34.1); DEK-NUP214 AML with inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2); GATA2, MECOM AML (megakaryoblastic) with t(1;22)(p13.3;q13.3); RBM15-MKL1 AML with BCR-ABL1 (provisional entity) AML with mutated NPM1 AML with biallelic mutations of CEBPA AML with mutated RUNX1 (provisional entity)
AML with myelodysplasia-related changes	
Therapy-related myeloid neoplasms	
AML, not otherwise specified (NOS)	AML with minimal differentiation AML without maturation AML with maturation Acute myelomonocytic leukemia Acute monoblastic/monocytic leukemia Acute erythroid leukemia Pure erythroid leukemia Acute megakaryoblastic leukemia Acute basophilic leukemia Acute panmyelosis with myelofibrosis
Myeloid sarcoma	
Myeloid proliferations related to Down syndrome	Transient abnormal myelopoiesis Myeloid leukemia associated with Down syndrome

Table 1.1: Overview of the 2016 classification of AML according to the WHO, focussing on genetic properties. Table adapted from [7] and [8].

1.1.3 Classifications

According to the current classification scheme published by the World Health Organization (WHO), AML can be further classified using a variety of morphological, immunophenotypical and genetic criteria, as well as clinical findings [7]. A brief overview of the structure of the 2016 WHO classification is given in Tab. 1.1.
While recent versions of the WHO classification increasingly rely on a genetic characterization of AML, the so-called French-American-British (FAB) classification is also still commonly used. It forms the basis for describing cases which do not fit into a genetically defined subtype and are subsumed under the category “AML, not otherwise specified” in the WHO classification (cf. Tab. 1.1 and 1.2.). First established in the 1970s [12] and later extended [13], that classification is primarily inspired by morphological criteria. A summary of the main classes included in the updated FAB classification is shown in Tab. 1.2.

1.1.4 Epidemiology

A detailed understanding of the origins of the genetic alterations leading to AML remains elusive today. However, a number of risk factors are known to be associated with an increased risk. These include environmental factors such as exposure to benzene, ionizing

FAB Class	Description	Comments
M0	AML with minimal differentiation	No Auer rods.
M1	AML without maturation	Rare Auer rods.
M2	AML with maturation	Auer rods possible. Most common type (30-40% of cases).
M3	Acute promyelocytic leukemia (APL)	"Faggott cells" with many Auer rods. M3v variant has bilobed nuclei in peripheral blood.
M4	Acute myelomonocytic leukemia (AMMoL)	Auer rods uncommon. M4eo variant has abnormal eosinophils.
M5	Acute monoblastic leukemia	No Auer rods. Further subclassification possible: M5a: Monoblasts predominant M5b: Monocytes predominant
M6	Acute erythroid leukemia	Presence of multinucleated erythroblasts and myeloblasts.
M7	Acute megakaryoblastic leukemia	Polymorphic blasts. May include cytoplasmic vacuolization.

Table 1.2: Overview of the FAB classification, primarily based on morphological properties. Note that apart from the M3 subclass, the FAB scheme is included in the subclassification of "AML, not otherwise specified" type of the 2016 WHO classification (cf. Tab. 1.1). Table compiled from Refs. [9, 10, 11].

radiation, alkylating agents and cigarette smoke, as well as non-modifiable risk factors such as old age and male sex [11, 14, 15]. Additionally, some chromosomal genetic disorders such as the Down, Klinefelter and Turner syndromes are associated with an elevated risk [11]. Furthermore, AML can also arise secondary to a previous hematipoietic disorder, usually Myelodysplastic Syndrome (MDS) or Myeloproliferative Neoplasm (MPN) [16]. Both AML and MDS can emerge as a complication of previous medical treatment, such as cytotoxic chemotherapy or radiation therapy.

1.1.5 Clinical presentation and treatment

On an epidemiologic level, AML represents the most common acute leukemia in adults, with an incidence of 3 – 4 per 100,000 persons per year [15]. Median age at diagnosis in large populations tends to be found in the range of 66 – 74 years [15, 17]. Most prominent clinical findings in AML patients include general symptoms such as weakness, fatigue and fever. Pallor, increased infection risk, easy bruising and increased risk of bleeding reflect hematopoietic derangement, leading to anemia, neutropenia and thrombopenia. Certain AML subgroups can exhibit more specific symptoms, such as infiltration of the gums by blasts in the FAB groups M4 and M5 [11].
Without treatment, AML has a very poor prognosis with survival times of days to few months [8]. It was only with the advent of chemotherapy from the 1950s onwards that this natural history could be significantly improved [18].
The current standard initial treatment still consists of a chemotherapy backbone with the

aim of inducing remission, i.e. reduction of blasts in the bone marrow under 5%, and a normalization of peripheral leukcocyte and thrombocyte counts. In order to achieve this aim, an induction treatment with cytarabine and an anthracycline according to the so-called 7+3 scheme remains a standard of care [19]. Eventually, a significant number of patients relapse and become non-responsive to further therapeutic intervention. As an addition or alternative to chemotherapy, hematopoietic stem cell transplant can be considered in suitable patients [20].
Overall, AML still today has a poor prognosis, and 5-year survival rates in adult patients can be as low as 10% [21]. A notable exception is Acute Promyelocytic Leukemia, whose treatment has dramatically improved thanks to the introduction of regimens based on all-trans retinoic acid (ATRA) and arsenic trioxide (ATO), with cure rates reaching well over 80% [22].

Advances in the understanding of the genetic landscape underlying the disease biology of AML have led to a considerable diversification in the identification of patient subgroups [19]. This has triggered the development of various novel therapeutic approaches, and an expansion of the range of available drugs whose success remains to be thoroughly evaluated [23]. The hope is to ultimately arrive at treatment strategies which use the detailed knowledge of disease mechanisms acquired in recent years in the framework targeted and personalized therapies [24, 25, 26].

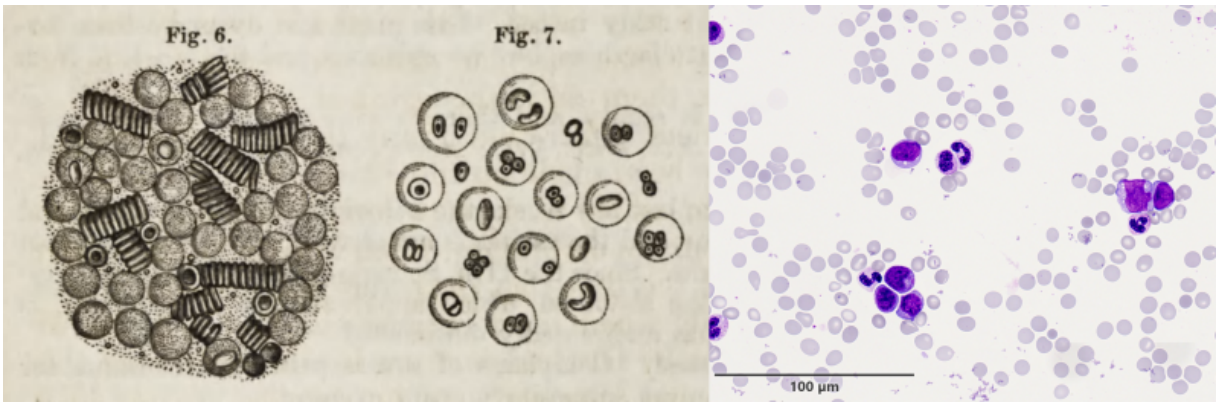


Figure 1.2: Historical and present-day leukocyte images.
Left: Drawings of leukocytes in a leukemia patient published by Bennett in 1851 [27], thought to be the first depictions of cell morphology in patient with that disease [18].
Right: Scan of a peripheral blood smear of a leukemia patient in Pappenheim stain, obtained with the scanner device used for the work presented in this thesis.

1.2 Diagnostic modalities

1.2.1 Light microscopy

The methods included in the diagnostic workup of AML cover a broad range of technological aspects, and have co-evolved with the increasingly detailed understanding and sub-classification of the disease. This necessitates use of a several diagnostic methods in order to arrive at a final diagnosis. Morphological examination of peripheral blood smears and bone marrow samples under a light microscope is by far the oldest method, which however remains a key step of the diagnostic algorithm [28]. As it is the main imaging method used in this work, its basic properties are briefly discussed here.
The technological roots of light microscopy, including use of water and oil immersion techniques, can be traced back to the pioneering works of Hooke [29] and van Leeuwenhoek [30] in the 17th century. Improvements in the quality of microscopes by Abbe and others enabled their use in the investigation of pathologic alterations in tissues and cells, which also lead to the first modern descriptions of leukemia [2, 1] (cf. Fig. 1.2).
Today, microscopic evaluation of cytologic specimina remains an important cornerstone in the diagnosis of AML, and high resolution represents a key requirement of diagnostic quality. In this context, an optical objective magnification of 63x to 100x is usually required [31]. Furthermore, oil immersion is commonly used. The primary effect of this method is a decrease in the minimum distance δ required for two points to be optically discernible, which is given by [32]

$$\delta = 0.61 \frac{\lambda}{NA}. \tag{1.1}$$

Here, λ is the wavelength of the illuminating light, and $NA = n \cdot \sin\alpha$ the so-called numerical aperture of the objective used, which in turn depends on the refraction index n of the surrounding medium and the geometric half-opening angle α of the optical system (cf. Fig. 1.3). Intuitively, the numerical aperture NA is a measure of the system's ability to focus incoming light. Higher values of NA correlate with a higher transversal resolution, i.e. the ability to separate points geometrically close in the transverse direction of the optical axis. Hence, the objective resolution can be increased by replacing the air around the objective ($n_{air} \approx 1.0$) with an optically denser medium with higher refractive index (cf. Fig. 1.3). Standard immersion oil used today typically has values of $n \approx 1.5$.
Another important optical parameter in the use of microscopes and scanners is the depth of field (DOF), i.e. the maximum distance from the focal plane at which an object can be simultaneously in focus. Using an optical detector whose smallest resolvable distance is given by e, the DOF of a microscope with magnification M is given by [32]

$$d_{DOF} = \frac{\lambda \cdot n}{NA^2} + \frac{n}{M \cdot NA} e. \tag{1.2}$$

For microscopes with high magnification and numerical aperture as used in cytomorphology, the depth of field predicted by eq. 1.2 can be well below 0.5 μm. Normally, a shallow DOF is compensated by focussing through the sample when examining one field of view

with an optical microscope. For scanners, that image larger regions of a coverslip consisting of many fields of view, this shallow focus depth means that small inhomogeneities of the coverslip surface can cause the image to defocus. This necessitates frequent re-focussing or elaborate focus tracking methods, which in turn tend to considerably slow the scanning process [33].

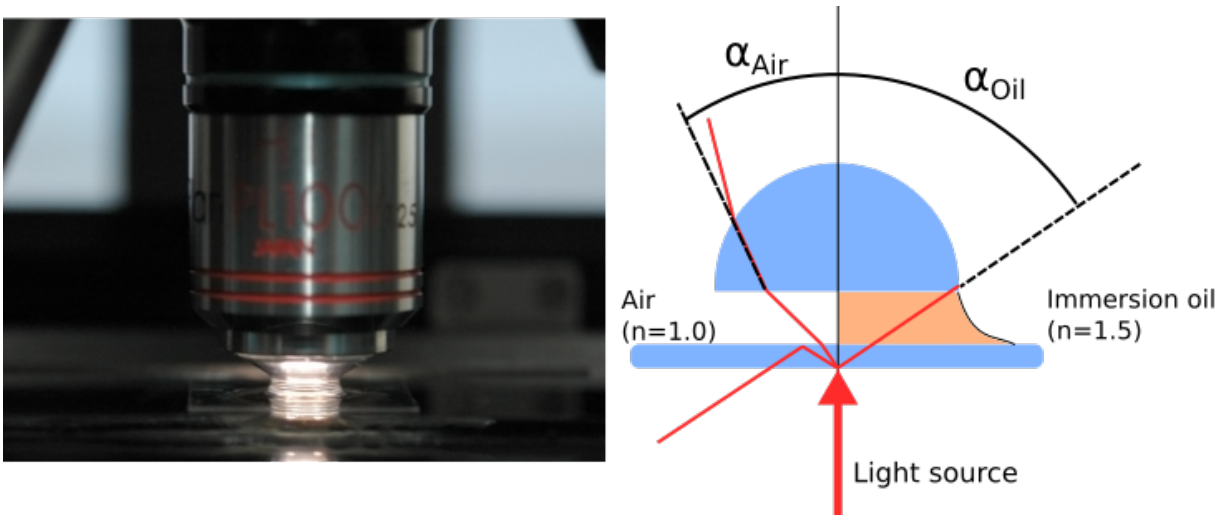


Figure 1.3: Oil immersion in light microscopy.
Left: Microscope with objective immersed in oil, as commonly used in cytomorphologic examination. Adapted from [34].
Right: Schematic depiction of optical path with air and immersion oil as surrounding medium. Use of optically dense immersion oil results in an increase of numerical aperture $NA = n \cdot \sin \alpha$, and hence resolution.

1.2.2 Microscopy sample preparation and staining

For sample preparation of peripheral blood smears, a droplet of capillary or venous blood anticoagulated using an ethylene diamine tetraacetic acid (EDTA) buffer is put at the edge of a ground cover glass, and is then smeared out using the technique shown in Fig. 1.4. In order to morphologically differentiate the cellular components of the smear, different common staining protocols are available [36]. In Europe, the most frequently used protocol is a combination of the May-Grünwald and Giemsa stains, which is also known as Pappenheim stain after its original developer [37]. Its main ingredients include eosin, methylene blue as well as azure A and B [36]. A very similar alternative is the Wright stain, which is more popular in the United States [38]. Both stains are panoptic, with eosinophilic, basophilic and neutrophilic components. While these standard panoptic chemical stains are the most frequently used stains for hematological cytomophology, many more staining methods exist for specific purposes, including e.g. the Heilmeyer stain for reticulocyte counting, or the Berlin Blue Iron stain for detection of trivalent iron [39].

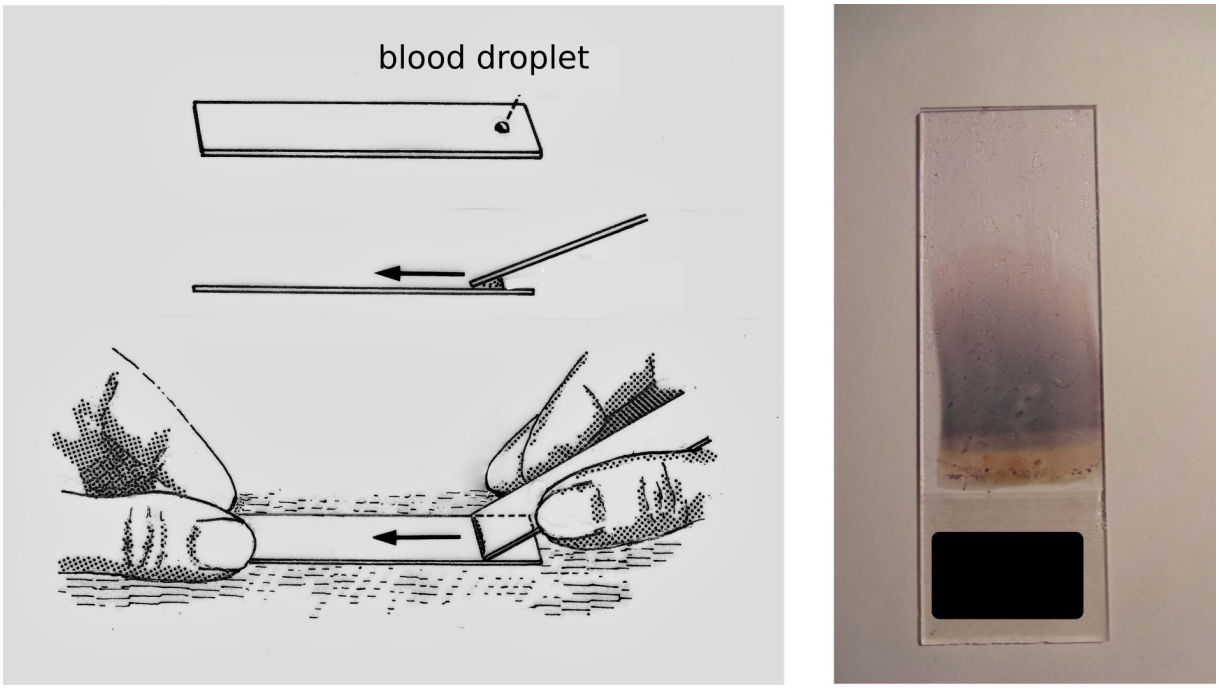


Figure 1.4: Preparation of peripheral blood smears.
Left: To produce a blood smear, a droplet of blood is placed on a coverslip and smeared out in the fashion schematically shown. Modified from [35].
Right: Macroscopic aspect of a blood smear sample prepared as shown on the left and stained using a Pappenheim protocol.

Another class of important stains relevant in the diagnosis of AML are cytochemical stains, whose behaviour relies on the use of cellular enzymes and that are relevant in the context of the FAB classification of AML [13]. The most important stains from this class are the myeloperoxidase, nonspecific esterase and PAS stains.
The focus of the present work is on cytomophology from standard stains, hence all samples included were stained using a routine Pappenheim protocol. A typical sample is shown in Fig. 1.4.

1.2.3 Other diagnostic methods

In the light of a more detailed understanding of the disease biology of AML, several other methods have rapidly become integral parts of the routine diagnostic workup of myeloid neoplasms. Most of these methods yield intrinsically more quantitative data, but are much more costly and technologically complex, and are therefore often used only after initial cytomorphologic examination. As this thesis is concerned with microscopic examination of blood smears, they are mentioned briefly here for completeness. In a routine setting, all methods are increasingly used concurrently in an integrated fashion, allowing to arrive at a more complete picture of the entity in question [40]. This includes molecular information

that bears direct relevance for the treatment and prognosis of the disease, such as the BCR-ABL or FLT3 markers [41].
A key aspect in the characterization and sub-classification of Leukemias is determination of the immunophenotype. This is typically done using flow cytometry, which allows sorting of cells according to parameters such as cell size, granularity or antigen pattern present in the cytoplasm or on the cell surface [9]. This allows quantitative allocation of cells to defined subclasses defined by characteristic markers (cf. Fig. 1.5). Depending on the suspected class, comprehensive consensus marker panels have been established [42].
Furthermore, genetic methods represent a large and increasingly important part of leukemia diagnostics. This category includes cytogenetics, i.e. interpretation of number and banding patterns of stained metaphase chromosomes (cf. Fig. 1.5), which can exhibit characteristic alterations in AML [20], and is essential to classifying the disease according to the current WHO classification (cf. Tab. 1.1). Genetic methods also comprise fluorescence in-situ hybridization (FISH), a technique that relies on the use of nucleic acid probes complementary to defined sequences on chromosomes that are tagged by fluorescent labels. Therefore, they allow targeting specific genetic alterations on a sub-microscopic scale, e.g. a rearrangement of the PML and $RARA$ genes in the framework of acute promyelocytic leukemia [43, 7].
Finally, molecular genetics plays an increasingly important part in the understanding and classification of AML. In the past few years, a rapidly expanding list of molecular biomarkers have been introduced, which allow for specific characterization of genetic alterations in diseased cell populations and hold considerable potential for the future diagnostic workup [40]. Furthermore, PCR-based techniques have become a very sensitive tool in monitoring and quantifying the disease kinetics of post-therapy AML through detection of minimal residual disease [40, 44].

1.3 Artificial Neural Networks for image classification

Models of computation that try to emulate the human thought process as "artificial intelligence" have a longstanding history that go back to the first modern concepts of computing machines in the mid-19th century [46, 47, 48]. Over time, these concepts have led to a large variety of approaches to define and concretely realise "artificially intelligent" algorithms. One line of thought derives from the idea to computationally emulate neural processing in the brain, leading to the concept of "artificial neural networks". Following the theory of learning based on neural plasticity developed by Hebb in the 1940s [49], a first artificial neural network, the perceptron, was developed by Rosenblatt in 1958 [50] as a pattern recognition algorithm. Work by Hubel, Wiesel and others on the neurophysiology of information processing in the visual cortex added biological plausibility to the use of artificial neural networks in recognising visual patterns [51]. However, computers available at the time of their first study made training of networks large and fast enough to be used for practical tasks difficult due to their limited performance and memory availability. Furthermore, results by Papert and Minsky outlined theoretical limitations of artificial neural networks [52], leading to a phase of reduced interest in the subject known as "AI

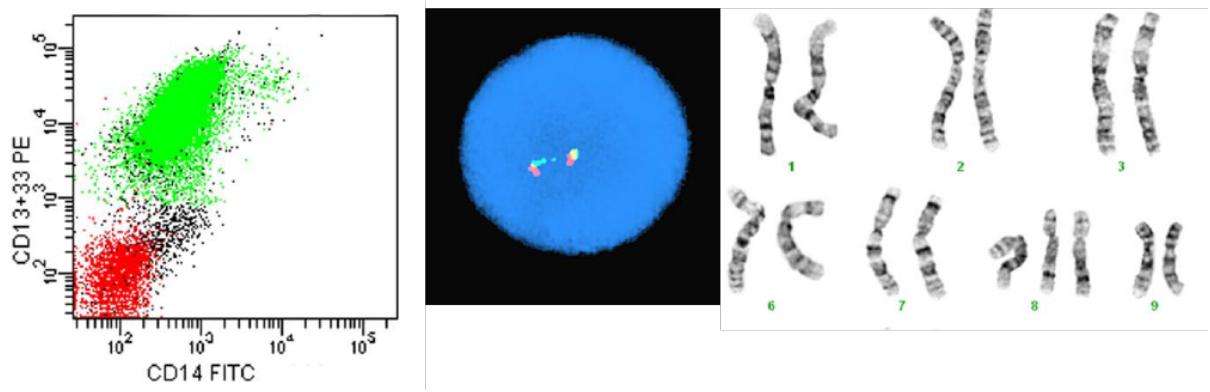


Figure 1.5: Common modalities in the diagnostic workup of AML.
Left: Dot plot resulting from flow cytometry of staining for the cell surface markers CD13 and or CD33 versus CD14. The myeloid population of interest is coloured green, and a small lymphoid population in red.
Middle: Dual fusion signal in a FISH study, in this case using an LSI MLL dual color DNA probe.
Middle: Typical cytogenetic study, revealing changes to chromosomal structure and an additional chromosome 8.
Images reproduced from Ref. [45]

winter" [53]. Nevertheless, the following years led to important advances, such as development of the *backpropagation* algorithm, which is important for efficiently training artificial neural networks [54, 55] and remains in widespread use.

While neurophysiological findings remain an inspiration for the development of artificial neural networks today, their practical design and training is typically independent of direct biological models. Neural networks used today consist of basic, abstract processing units, which produce an output by applying a usually nonlinear activation function to several input values. In this regard, their structure is inspired by the function of biological neurons, which receive and integrate synaptic activations as inputs that influence their firing behaviour and produce an output signal that is transduced via the axon. However, the details in the implementation of abstract neurons differ considerably and lack a biological correlate [55] (cf. Fig. 1.6). Output and input channels of individual neurons are connected to form an abstract neural network. Training such an abstract neural network hence amounts to finding a set of weights in the neural connections that lead to a desired collective behaviour of the network. In particular, sequential stacking of multiple layers of neurons between input and output layer, forming so-called hidden layers, has turned out to be an important aspect in the development of successful applications. Networks with a large number of hidden layers are called *deep* networks. Increasing the depth (i.e., number of consecutive layers) has been a major driver in the development of more successful CNNs [56]. Machine learning methods that rely on the use of deep artificial neural

networks are often subsumed under the term “deep learning”.

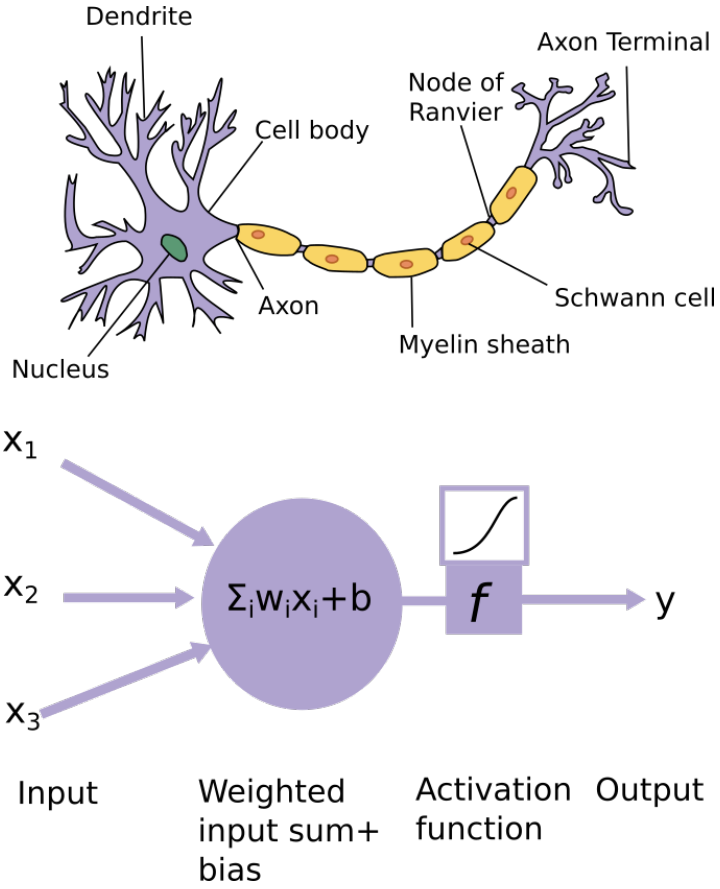


Figure 1.6: Analogy of biological neurons (upper schematic) and neurons used in artificial neural networks (lower schematic): If a weighted sum of input values reaches a certain threshold (modelled by the bias b), a nonlinear activation function produces a high output in analogy to the “firing” of a biological neuron. Learning occurs by modification of the input weights to produce a desired behaviour of the overall neural network.
Upper schematic reproduced from Ref. [57]

Consistently, the present success of deep neural networks in a growing number of practical tasks was partly enabled by advances in available training strategies for deep networks from around 2006 on [58, 59]. Furthermore, the wide availability of graphics processing units (GPUs), which are optimized towards fast matrix and vector operations important in neural network training, has helped the popularity of deep neural networks [56]. Image classification has emerged as one of the most successful applications of deep neural networks since 2012, when a contribution by Krizhevsky and coworkers that used a deep convolutional neural network (CNN) won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) by a large margin, beating other approaches that relied on a hand-crafted feature extraction strategy (cf. Fig. 1.7) [60]. Some of the hidden layers of CNNs

are convolutional layers, which leads to a shift invariant classification behaviour [55]. In addition to this success in natural image recognition tasks, deep neural networks showed similar success in medical imaging, such as the detection of mitoses in histological images [61]. Since then, a number of key medical image classification tasks have been addressed using CNNs, including classification of skin lesions [62], retinal fundus photographs [63, 64] and mammographic images [65].
In the intervening years, deep neural network-based models have been consistently dominating the list of best-performing image classification algorithms, and have dramatically increased the overall success rate of automated image classification. This development is consistently witnessed e.g. by the results of the ILSVRC competition (cf. Fig. 1.7).

1.4 Scope of the project

In this thesis, a project is described that aims to develop a diagnostic support system for cytomorphologic classification based on the use of state-of the art Convolutional Neural Networks (CNNs). As is common for neural network-based approaches, successful training and evaluation of a neural network for classification tasks typically requires a large amount of annotated data. In this work, a digital leukocyte image database is compiled based on the peripheral blood smears of 100 AML patients and 100 persons without morphological signs of malignancy. This database is then used for both training and evaluation of several neural networks. The quality of the underlying image dataset is extensively evaluated using re-annotation. Furthermore, the networks trained are analysed as to which parts of the classified images are relevant in producing the classification result.
As a result of the scanning and labelling process, a database of more than 18,000 single-cell leukocyte images is set up in this work, which is the largest publicly available morphological image database available to date. Furthermore, testing of the networks trained using this dataset shows that these algorithms reach human-level performance at the task of morphologically classifying individual leukocytes, hence making the power of deep learning models accessible to an important area of hematological diagnosis.

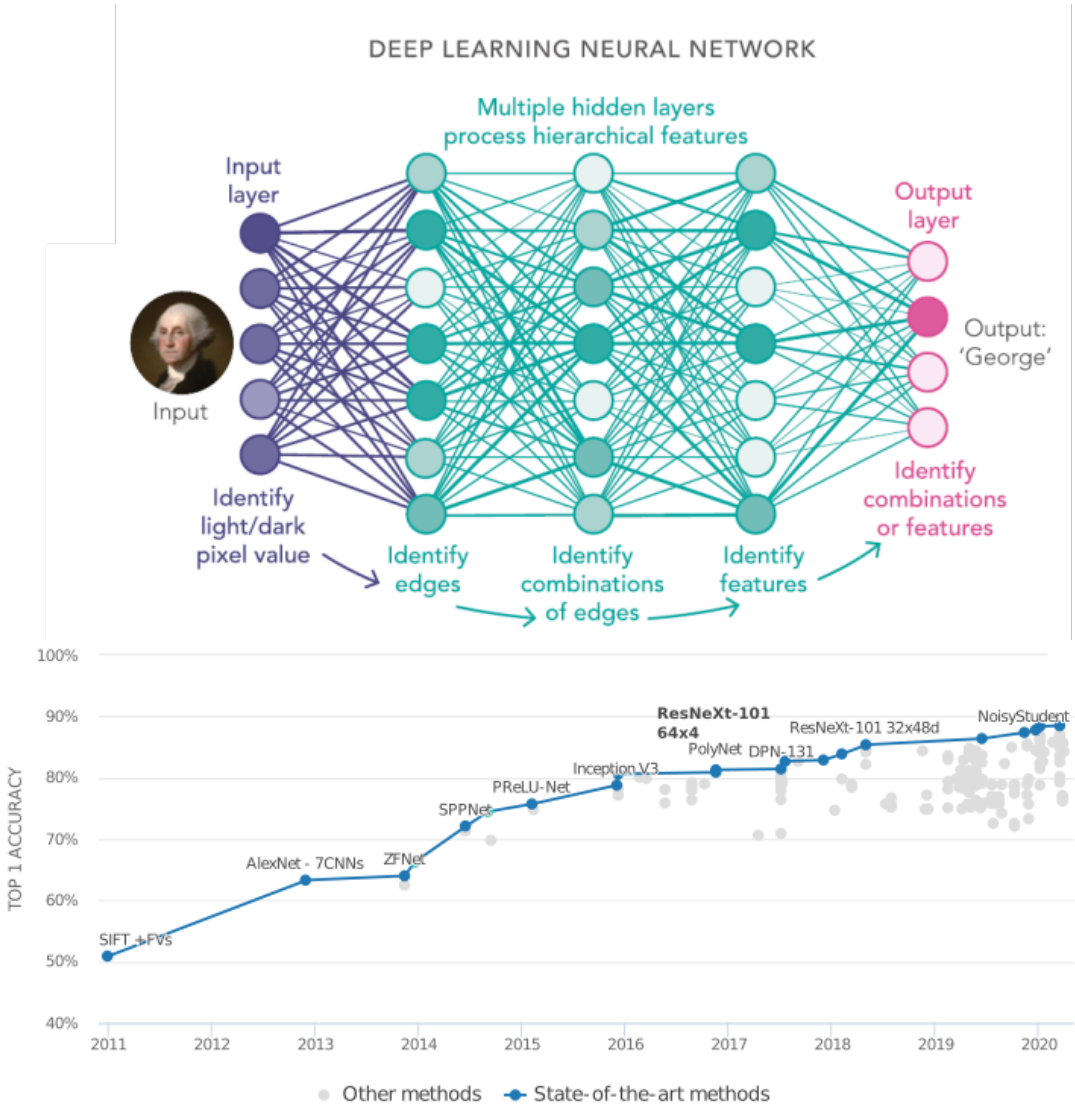


Figure 1.7: Structure and performance of modern CNNs:
Upper panel: Typical structure of a modern deep neural network. Several hidden layers lie between the input and output nodes, and contain increasingly abstract representations of the input signal. In the final layer, this leads to a classification (in this case, the first name of the pictured person). Figure reproduced from Ref. [66].
Lower panel: Classification accuracy of algorithms submitted to the ILSVRC contest since 2011. Note the steep increase in accuracy in 2012 due to AlexNet by Krizhevsky *et al.* [60]. First submission of the ResNeXt model used in this work is shown in bold. Figure modified from Ref. [67])

Chapter 2

Materials and Methods

A prerequisite for the use of computational analysis methods for image-based single-cell classification is high-quality sample digitization using appropriate microscopic imaging devices. As the gold standard for morphological classification is defined by the trained human cytologist, the next step in developing an automated classification system consists of annotation of the acquired material. The result is a comprehensive dataset of microscopic images annotated on the single-cell level that can be used to train and evaluate neural network models.
This chapter gives an characterization of the cohort of patients whose blood smears were included into the study. Then, the digitization and annotation process is described. Finally, the network structures and main training and analysis tools are introduced.

2.1 Cohort selection and properties

For much of classical statistical modelling, the size of the dataset included in the development of a model plays an important part in the ability to derive robust statements. In a data-driven method such as deep learning, the aim is to train a model that can be successfully applied to previously unseen data, i.e. a model which possesses favourable generalization properties. Overall, the intuitive expectation is that models trained on larger training datasets tend to generalise better [55, 68]. At the same time, it has been shown that fewer but carefully chosen training samples covering the information necessary to build models may suffice for certain networks and datasets [69]. Presently, no general, precise way of estimating the number of samples needed for training a neural network with a desired performance is known. For single cell classification, it might however be expected that the training data should at least cover the range of observed morphologies with several independent cases. While in this work, models are trained to classify single cells rather than the entire blood smear of a patient, collecting cell images from several patients for each class is expected to aid generalisability. At the same time, certain cell populations are rare, even in a large patient cohort. Ultimately, the number of patients and cells included reflects a tradeoff between scan and annotation time against the aim of creating a dataset

that covers the range of possible appearances morphological classes considered reasonably well.
For the present work, peripheral blood smears from 100 patients diagnosed with AML at the Laboratory of Leukemia Diagnostics at LMU Klinikum were selected. The patient group covered most morphological classes of the FAB scheme (cf. Tab. 1.2). The distribution of FAB classes in the AML patient cohort is shown in Fig. 2.1.

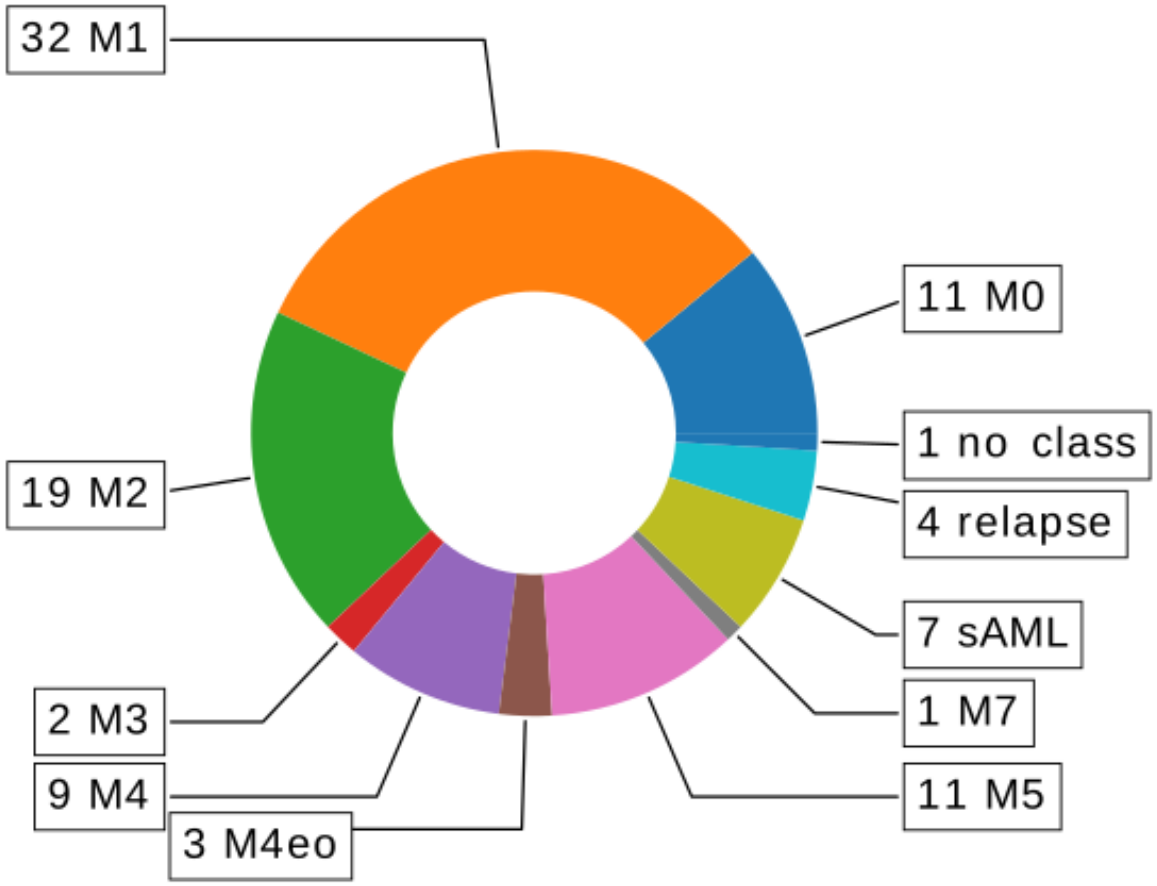


Figure 2.1: FAB class distribution of patients included in the AML cohort.
Figure adapted from Ref. [70].

To avoid bias for the cytomorphology in blood smears of AML patients, 100 control blood smears were also included. While the control smears were taken from patients at the LMU Klinikum rather than healthy controls, they were found not to exhibit morphological signs of malignancy before inclusion into the study. All blood smears included were evaluated within the routine workflow of LMU Klinikum between 2014 and 2017. The study set-up was reviewed by the ethics committee of the LMU medical faculty, and consent was obtained under reference number 17–349.
A distribution of age and gender of the AML patients and controls included in the study

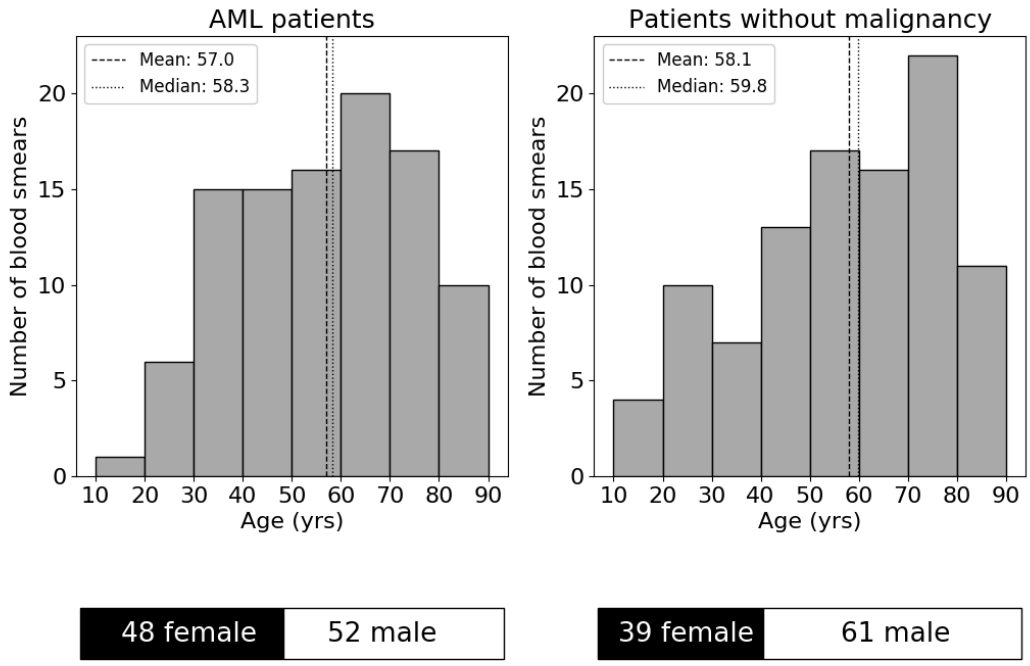


Figure 2.2: Distribution of age and sex in the AML patient and control group.
Upper panels: Age distributions, showing a similar median age for AML and control patients with 58.3 and 59.8 years respectively.
Lower panels: Sex distributions in both cohorts. Figure reproduced from Ref. [70].

is given in Fig. 2.2. Age properties of the AML and control groups were comparable, with a median age of 58.3 for the AML group and 59.8 years for the control patient group. The AML group showed a female/male ratio of 48/52, and the control group of 39/61. An outline of the data acquisition and processing workflow is shown schematically in Fig. 2.3.

2.2 Digitization process

Morphological evaluation of blood smears for AML diagnostics is done by primarily evaluating leukocyte cytomorphology. Evaluation is normally restricted to the so-called monolayer area of the blood smear, which is the region in which single blood cells lie densely, but without overlapping [5]. Digitization of a large enough part of the monolayer region is therefore sufficient for annotation of cells from a blood smear.
Reduction of the area to be scanned is important, as the scan parameters required for cytomorphology, i.e., 100-fold magnification and immersion oil use, lead to much higher scan times than in the case of histopathology, where 40-fold magnification is more frequently

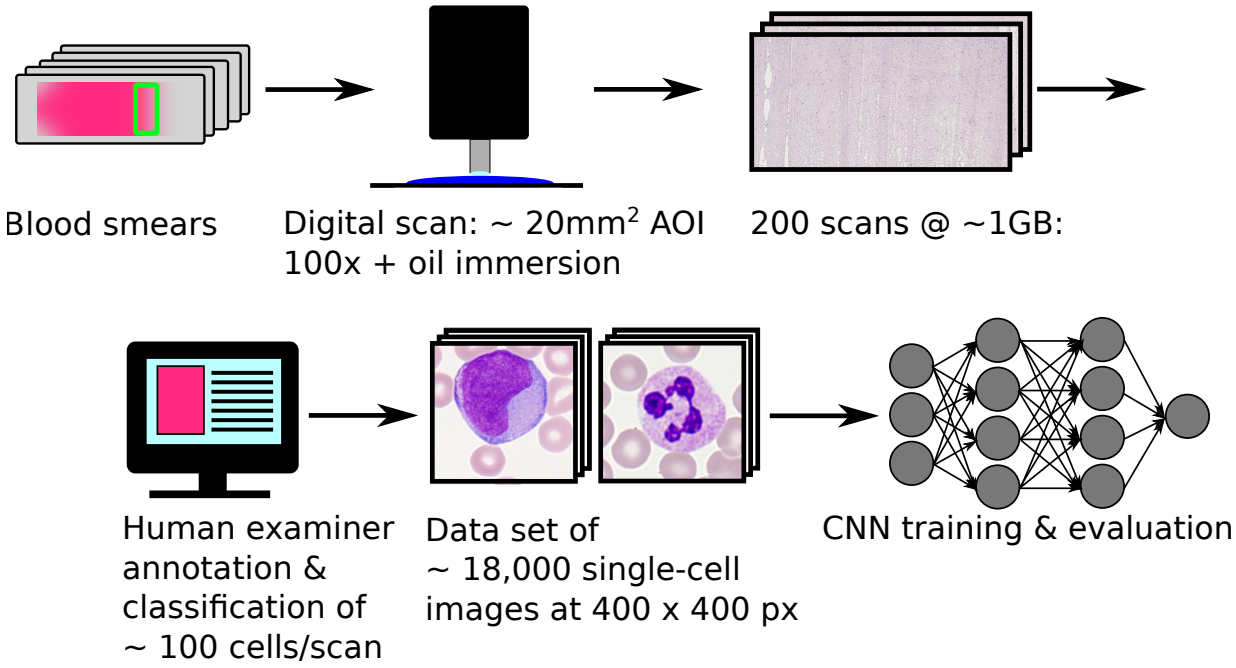


Figure 2.3: Schematic depiction of the workflow used in this study. In order to estimate intra- and inter-rater variability of annotation, single-cell patches were re-annotated up to two times after initial annotation from the AOI scan.
Figure reproduced from Ref. [71].

used [72, 73]. This effect is due not only to the increased number of fields of view (FOVs) to be imaged by the scanner, but also due to the necessity to refocus more frequently due to a shallow depth of field, as presented in Sec. 1.2.1. Additionally, high-resolution scans of large areas of interest (AOIs) can lead to prohibitively large file sizes, providing further motivation for scanning only relevant areas of the blood smear.
All slides included in the present study were digitized using the M8 digital microscope-scanner manufactured by Precipoint GmbH, Freising, Germany (cf. Fig. 2.5). In the scanning process, a low-magnification overview image of the smear was first produced, from which an AOI of approximately 20 mm^2 was manually selected in the monolayer region and then scanned at 100-fold magnification. Immersion oil was applied manually. Scan times depended on a variety of hardware settings, but normally ranged between 30 and 60 minutes per AOI. The scanner produced files in the vendor-specific .vmic format, which typically leads to file sizes of approximately 1 GB.

2.3 Annotation

As the classifiers based on convolutional neural networks (CNNs) in this work are trained and tested using expert-annotated data, size and quality of data annotation are key to classifier validity. In some use cases of CNN-bases classifiers, the ground truth of an image

annotation can be provided by an underlying gold-standard method. For example, the dignity of photographic images of skin lesions can be decided by histopathological examination [62]. In the case of leukocyte cytomorphology however, no obvious independent parameter exists which could definitively determine the morphological class of a given cell. For this reason, ground-truth image labels have to be provided by a human examiner.
In order to annotate the image data in a standardized way that allows estimation of intra- and inter-examiner variability, up to three annotations of individual leukocytes were performed. A first annotator used the entire AOI that was scanned from the monolayer region of the blood smear as described in Sec. 2.2. The examiner was asked to proceed as in the normal case of evaluating a blood smear as far as possible, and flag approximately 100 leukocytes per smear in the scanned AOI, and classify them into the morphological classification scheme shown in Fig. 2.4. The classification scheme is based on the scheme used in clinical routine, and at this stage included several subcategories that are characteristic for some subtypes of AML, such as faggot cells and bilobed promyelocyted for AML M3v [9]. During annotation, the examiner used a custom-written deep-zoom viewer described in Appendix A (cf. Fig. 2.5). The first annotator could access the whole scanned AOI and compare different cells.
Based on the first annotation, single-cell image patches of size 400 x 400 pixels around the positions flagged by the annotator were extracted from the AOI. When the point flagged by the first annotator lay closer than 200 pixels from the edge of the scanned AOI, the part of the patch outside the AOI was filled with transparent pixels, i.e. pixels with a zero alpha value. From the dataset obtained in this way, patches containing more than one leukocyte were removed to ensure that labels of single-cell patches were unique. Overall, this yielded a set of 18,365 single-cell images.
In order to estimate interrater variability of leukocyte annotation, a second, independent cytologist was asked to re-annotate a subsample of 1,905 single-cell images which contained all morphological classes of our classification scheme. During re-annotation sessions, the re-annotator only had access to single-cell patches, and, unlike the first examiner, could not access the whole AOI scan. This setup was chosen in order to mimick the single-cell classification task of the neural network. For an estimate of intra-rater variability of single-cell classification results, re-annotation of the same subsample was repeated 11 months after the first re-annotation by the same cytologist.
The results of the first annotation are expected to be of higher consistency, as the first examiner has access to the whole AOI, potentially providing contextual information unavailable to the second examiner when performing the re-annotations based on single-cell images. For this reason, results of the first annotation are treated gold-standard labels against which results of re-annotation and algorithmic classification will be compared.

2.3.1 Data augmentation

As morphological types of leukocytes occur with different frequencies on the blood smears digitised, the distribution of cell types in the single-cell database is intrinsically imbalanced. For example, segmented neutrophils are expected to be the most frequently en-

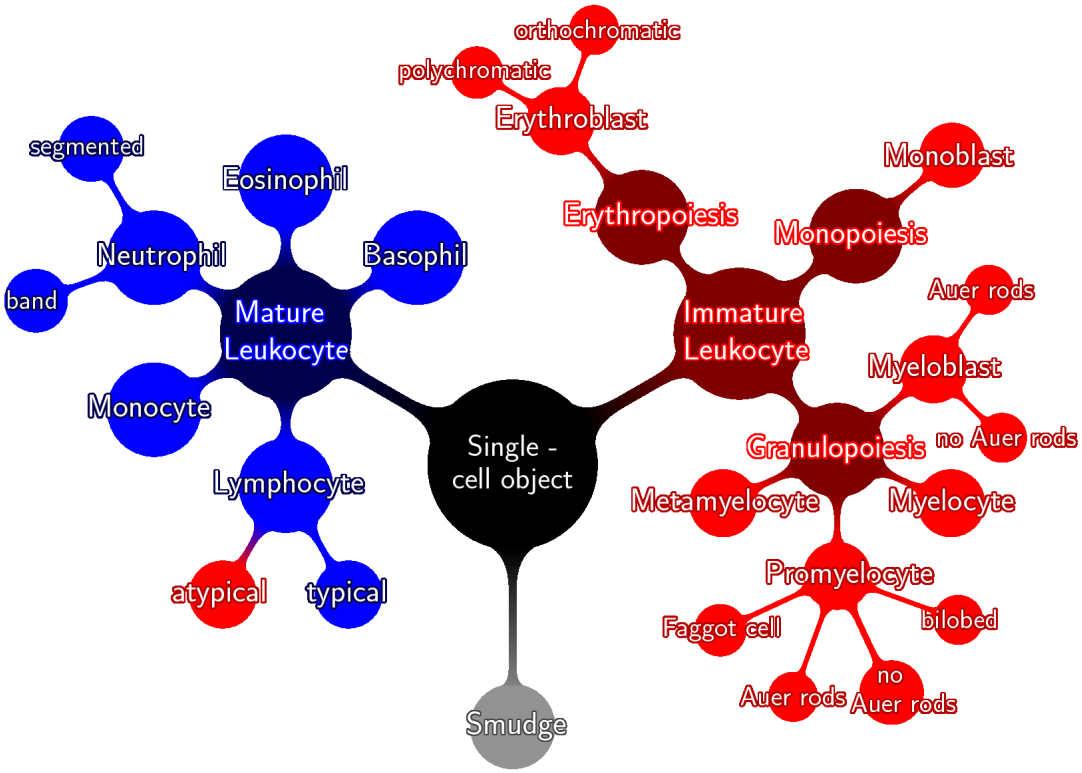


Figure 2.4: Taxonomy used for gold-standard annotation. The 19 classes corresponding to the leaves of the scheme are derived from the routinely used scheme [9], with some additional divisions to enable subclassification of promyelocytes and myeloblasts according to the presence of AUER rods and presence of a bilobed nucleus, and subclassify erythroblasts according to their staining behaviour.
Figure modified from Ref. [71].

countered leukocyte type under physiological conditions. In contrast, some cell types, such as monoblasts, are relatively infrequent and only few cases were classified in the gold standard annotation used in this work. Class imbalance in the training data can lead to poor generalization properties of the resulting model [55]. One of the common strategies to counteract this effect is data augmentation, which is based on the idea of upsampling the minority class by producing additional, "artificial" data through transformations of existing data. In the context of image classification, this strategy has been shown to increase model quality in many cases, even in the absence of class imbalance [55, 74].
For the present study, minority classes in the training data were augmented using horizontal and vertical flips, as well as random rotations in the range of $0° - 359°$. Scaling operations

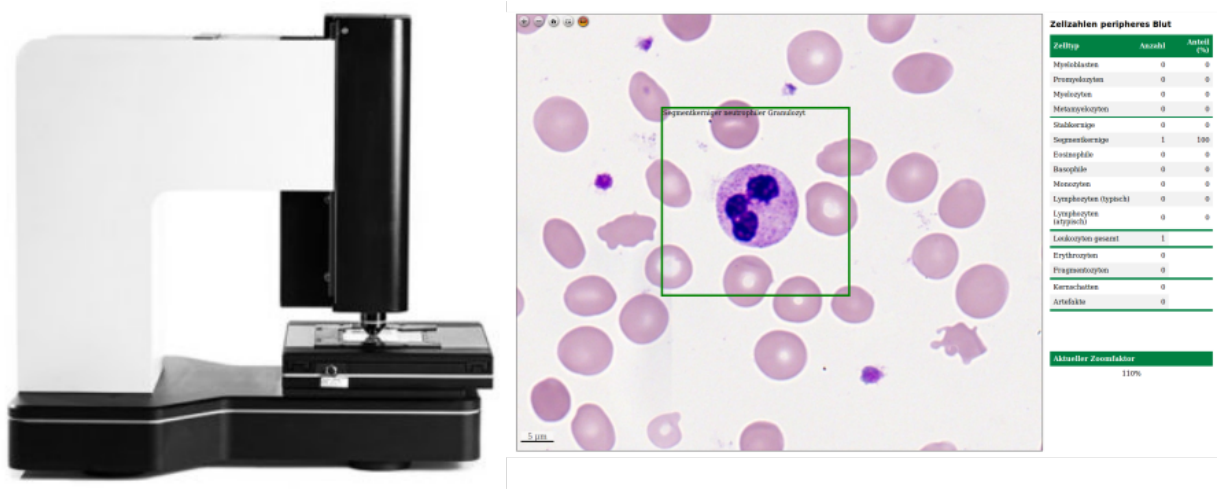


Figure 2.5: Equipment for blood-smear digitization and annotation. Left panel: The M8 microscope/scanner used for digitization of all slides included. Right: Custom-written annotation tool for annotation of single cells during the gold-standard annotation, as described in more detail in Appendix A.

were not used, in order to maintain cell size as a morphological criterion. Examples of the image augmentations used are shown in Fig. 2.6. The augmented dataset was saved and used for training all networks. Training data augmentation was performed after splitting the data into a test and a training set, in order to avoid contamination of the test set with augmented data from the training set. No augmentation was performed on the test set.

2.4 Computational methods

2.4.1 Hardware and software tools

Over the past few years, a number of frameworks have been developed for deep learning applications that allow a standardized setup, training and evaluation of neural networks [75], including the Microsoft Cognitive Toolkit (CNTK) [76], Theano [77] and Tensorflow [78]. These framework enable fast implementation and training of neural networks while encapsulating the underlying extensive matrix operations, making neural network development faster and less error-prone. Execution of code on GPUs is also supported, which is key to significantly speeding up neural network computations, which heavily rely on the use of matrix algebra. Throughout this work, the Keras library [79] encapsulating Tensorflow for the Python programming language was used for network implementation. All network training and evaluation for this work was performed on Nvidia GeForce GTX Titan X GPUs.

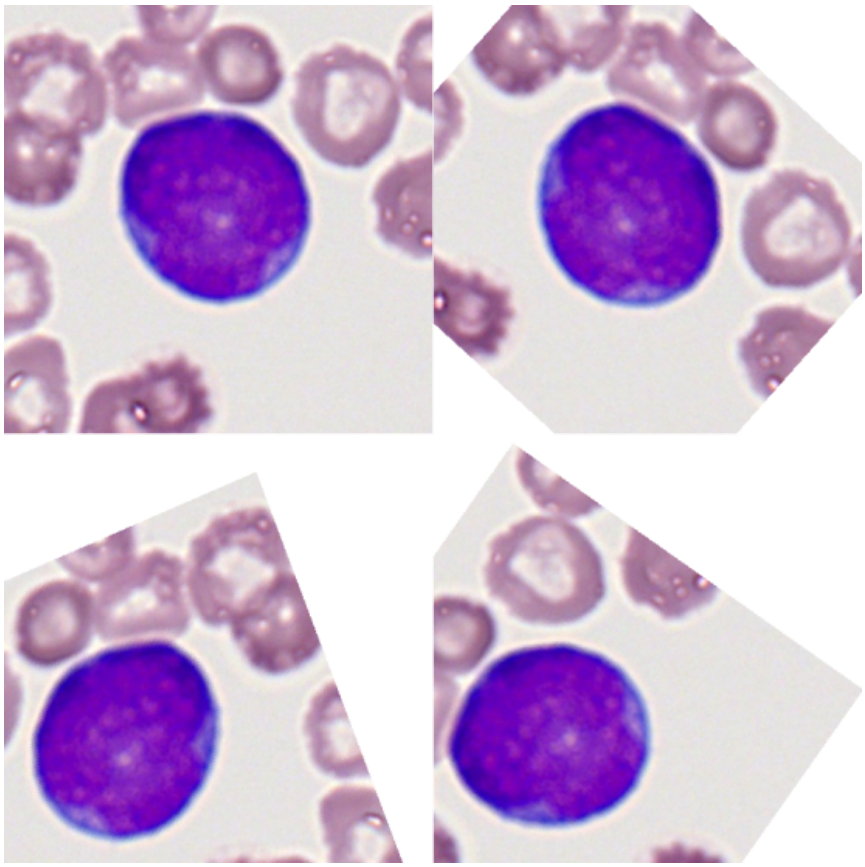


Figure 2.6: In order to counter class imbalance in the single-cell image dataset, data augmentation techniques were used in order to upsample underrepresented classes. Specifically, flips and random rotations were performed. Here, three images generated from the original single-cell patch annotated as a myeloblast (upper left panel) are shown.

2.4.2 Network architectures

In the context of image classification tasks, CNNs with a large variety of different network architectures and layer designs can be used [55, 80]. Choosing the best network structure in general and optimizing its hyperparameters in particular is generally a complex and resource-intensive problem [81]. The approach taken in this work is to use network structures that have proven successful in the context of natural image classification, and adapt them for the single-cell classification problem studied here. Systematic optimization of CNN hyperparameters is generally a complex and computationally costly problem [82], and was not attempted in the context of this work, which is a reasonable strategy for two reasons. Firstly, a number of highly optimised networks exist that have been used with increasing success for natural image classification in recent years [67]. As technically, cytomorphological classification fulfills a comparable task, these networks offer a good starting point. Secondly, the size of the image dataset studied in the context of the present project

is relatively small with approximately 18,000 images compared e.g. to ImageNet [83], the database used in the ILSVRC competition, which contains over 15 million labelled images [60]. Due to small dataset size, random effects due to sample noise are expected to be potentially more important than the gains of an improved network structure. In order to estimate sample noise, results in this work are stated as means and standard deviation obtained using k-fold cross-validation, as discussed in Sec. 2.4.3. In this work, a sequential network inspired by the VGG structure introduced by Simonyan and Zisserman of the Visual Geometry Group (University of Oxford), and the ResNeXt scheme [84] introduced by Xie *et al.* were used, and are described in the following.

Sequential network

This CNN follows the design philosophy of the VGG network [85], which was introduced in 2015 as an improvement of the seminal work by Krizhevsky and co-workers [60]. The network is built from stacked layers, which is termed a sequential model in the Keras environment. It contains four consecutive building blocks, which each consist of two 2d-convolutional layers, followed by a max-pooling layer. The final two layers are dense layers, after which the output is produced. After the first 2d-convolutional layer in each of the four building blocks, a so-called batch-normalization layer was introduced, which is used to increase the training stability of the neural network [86]. Activation layers following the 2d-convolutions use the so-called rectified linear unit ("ReLU") activation function defined as

$$f_{\mathrm{ReLU}}(x) = \max(0, x), \tag{2.1}$$

which is a commonly used choice for a non-saturating activation function, showing better training performance than other options in the context of image classification [60]. Kernel sizes of the 2d-convolutional layers were chosen as (3,3) in the first two blocks, and (6,6) in the second blocks, and are hence fairly small as in the VGG network [85]. A schematic overview of structure and parameters used for the sequential model in this work is shown in Fig. 2.7. Overall, the sequential network set up in this way contains 433,656 parameters, of which 433,224 are trainable. Networks with a structure similar to the sequential network used here represented the state of the art of image classification around 2015, with many improvements suggested since.

ResNeXt

After the initial success of CNNs in the realm of natural image classification in 2012, a large variety of more complex, refined networks have been developed [87]. One of the key methods to outperform initial sequential models of the type described in the previous section are so-called residual networks (ResNets). The reason for their initial development was the observation that networks of increasing depth turned out to yield higher training errors, an effect known as the degradation problem [88, 89]. As a solution to this problem, He *et al.* proposed the ResNet structure, which won the ILSVRC classification task in 2015 [89]. The basic structural idea of ResNets is the introduction of skip connections

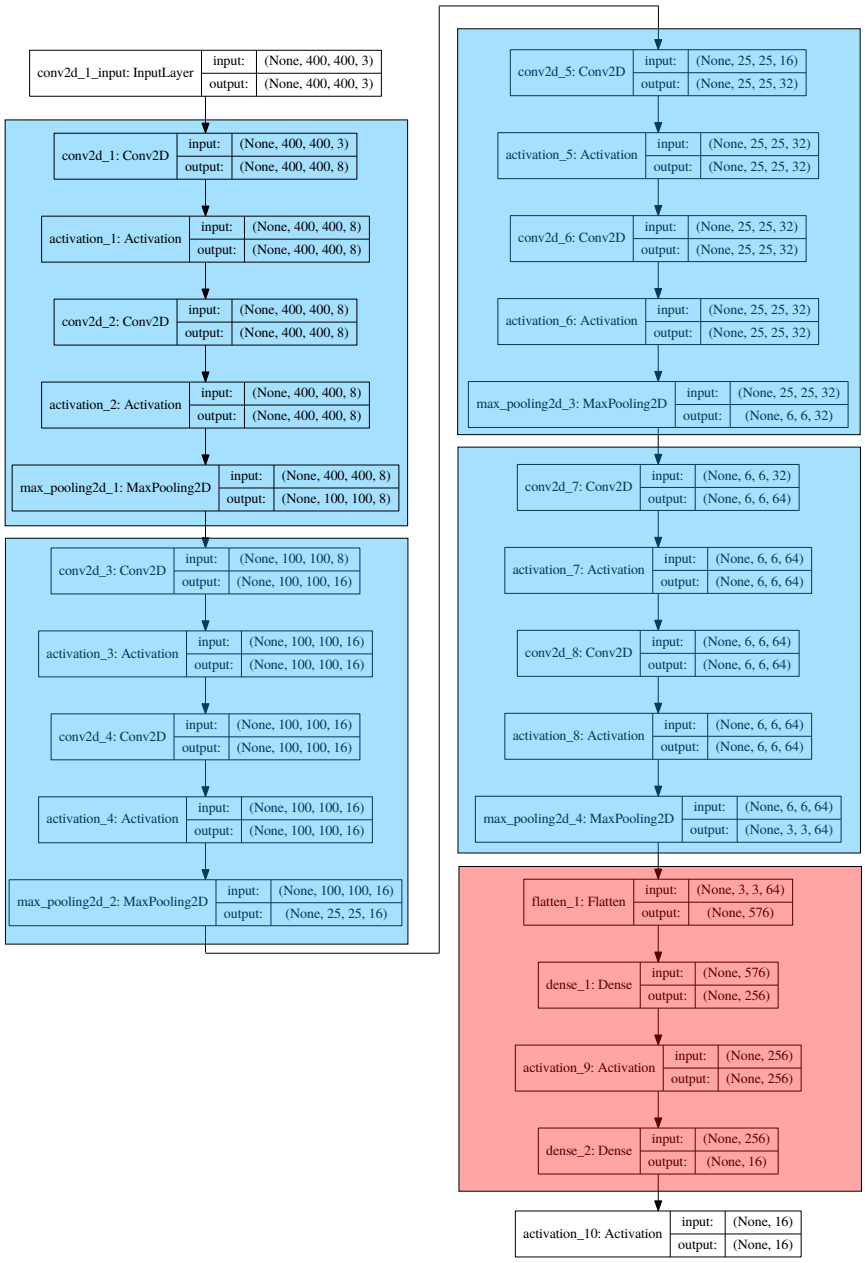


Figure 2.7: Detailed structure of the sequential model used in this work. In parallel to the seminal work published by Krizhevsky *et al.* in 2012 [60], the network contains several stacked convolutional layers with intervening max-pooling lavers. The convolutional layers are followed by two dense layers, which then produce the network output. Convolution parameters and network depth are inspired by the VGG network [85]. Analogous convolutional blocks are shaded blue, and the blocks containing the dense layers red.
Figure modified from Ref. [71].

that pass on input data via a shortcut consisting of an identity function, as is shown schematically in Fig. 2.8. Residual units hence split the transformations taking place in a layer into the form $\mathcal{H}(\mathbf{x}) = \mathbf{x} + \mathcal{F}(\mathbf{x})$, i.e. the sum of an identity function and a residual $\mathcal{F}$. Addition of layers in this structure can be trivially obtained by setting $\mathcal{F} = 0$, which might be expected to be easier to learn than building up an identity function from an additional stack of nonlinear layers [89].
After the success of ResNet, ideas have been proposed to further improve on that network in recent years, including Inception-ResNet [90], Wide Residual Networks [91] and ResNeXt [84], which achieved a second place in the classification task of ILSVRC 2016. The ResNeXt structure is used in this work as an example of an advanced, highly optimised network structure. The basic idea of ResNeXt as an improvement over the original ResNet structure is the decomposition of the residual function $\mathcal{F}(\mathbf{x})$ into aggregate transformations of the structure [84]

$$\mathcal{F}(\mathbf{x}) = \sum_{i=1}^{C} \mathcal{T}_i(\mathbf{x}), \tag{2.2}$$

where the $\mathcal{T}_i$ are functions of the same topology. The parameter C is called the cardinality of the ResNeXt block, and represents the number of parallel transformations in the block. A value $C = 32$ is used in Ref. [84], and maintained throughout this work. Overall, the core of ResNeXt consists of 16 stacked blocks of the kind described here [84]. The structure of both the ResNet and the ResNeXt blocks is shown in Fig. 2.8. In this work, an implementation of ResNeXt for Keras was used [92]. It is an attractive feature of the ResNeXt scheme as used in the context of the present work that it does not require fine-tuning of model hyperparameters. Rather, apart from adapting the input and output channels of ResNeXt, the hyperparameters used in Ref. [84] were maintained. As is expected for ResNeXt [87], the resulting model is big compared to the sequential model presented in the previous section, with an overall of 23,115,024 parameters, out of which 23,046,800 are trainable, leading to a relatively high computational cost.

2.4.3 Network training and evaluation

Following the strategy usually taken in machine learning, data was split into a training set and a test set at a ratio of approximately 80% to 20% [55]. All classes were split individually approximately according to that ratio, amounting to a so-called stratified split, in order to populate training and test sets with equal relative class compositions. After the split, only the training set was augmented as described in Sec. 2.3.1, and used to train the network, hence avoiding contamination of the test set with data related to the training set via an augmentation procedure.
As some classes contain only a relatively small number of single-cell images, the corresponding test sets for the classes involved can be small. For this reason, the test error estimate in particular is expected to be fraught with significant statistical uncertainty. In order to counteract this problem, the strategy of k-fold cross-validation was developed [55]. It consists of splitting the total dataset into k folds in a stratified way, and then training k

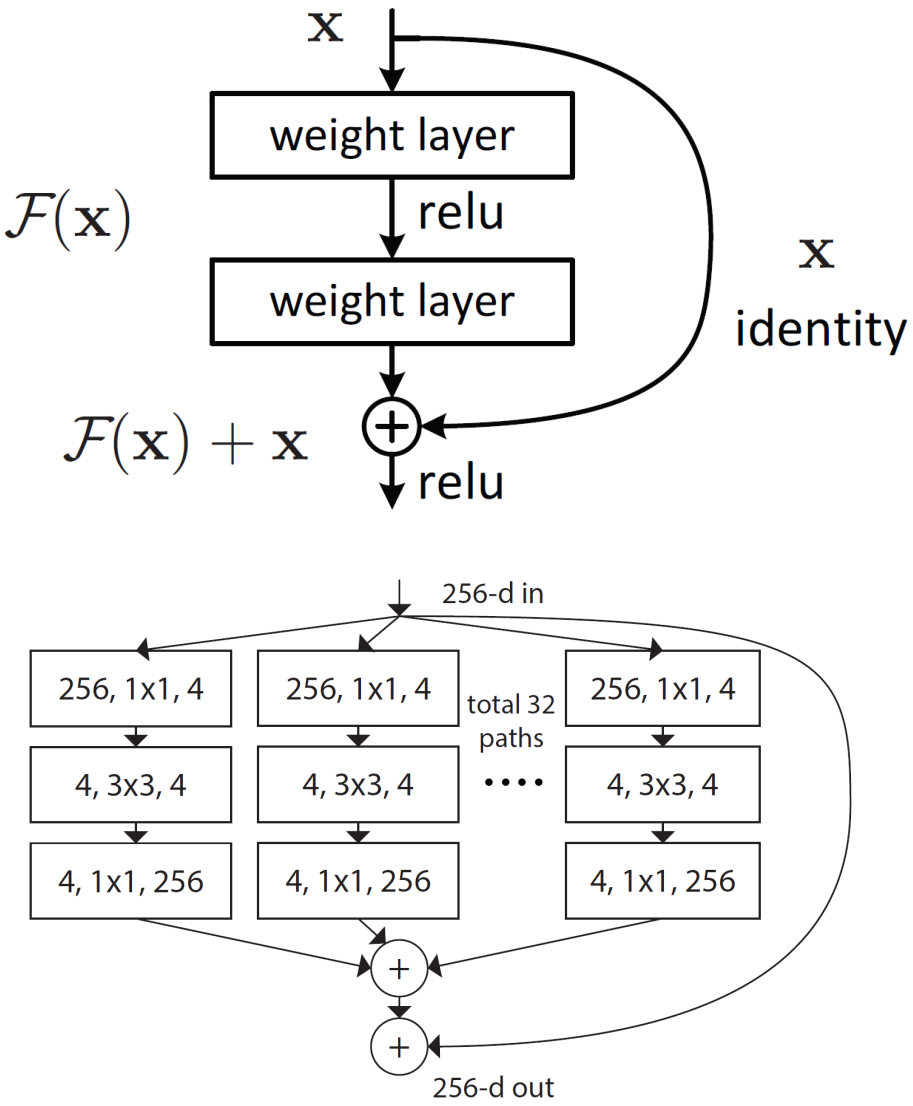


Figure 2.8: Schematic structure of blocks in ResNet and ResNeXt:
Upper panel: Basic building block of the ResNet, showing a decomposition of the network transformation into an identity and a residual function $\mathcal{F}$. Figure reproduced from Ref. [89].
Lower panel: Basic structure of the ResNeXt blocks, further dividing the residual function into C different parallel transformations.
Figure reproduced from Ref. [84].

models, in which one different fold is used for testing and the remaining folds for validation in each case (cf. Fig. 2.9). In the present work, this strategy was followed with $k = 5$, and results are given as mean $\pm$ standard deviation whenever possible, allowing an estimate of the statistical uncertainty of model evaluation introduced by the train-test split.
Supervised training is usually performed by feeding training data through the network,

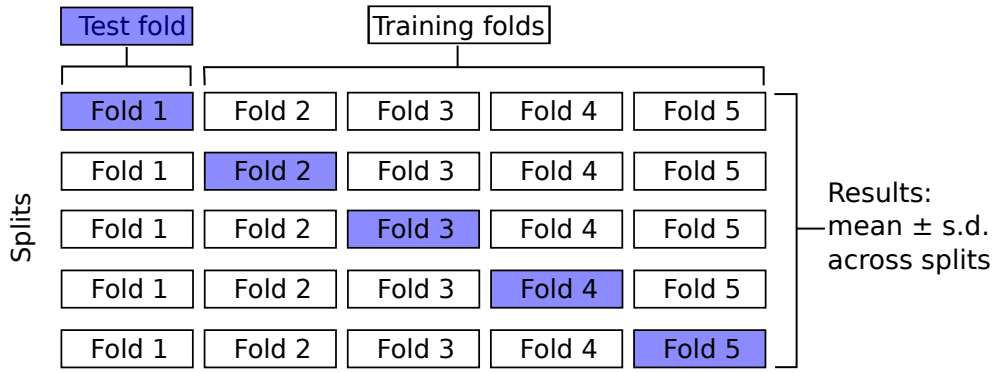


Figure 2.9: Schematic depiction of 5-fold cross-validation as used in this work. The overall data is split into 5 folds in a stratified way. Based on the splits, 5 networks are trained where each one has a different test set, and a modified training set. Results are then given as mean ± standard deviation across all 5 splits.

generating a prediction. A loss function, which measures the difference between network prediction and ground truth, is then calculated based on this prediction. Minimizing this loss function, i.e. bringing the model as close as possible to the ground truth, is the aim of training. Model parameters are usually initiated in a random fashion, and then iteratively refined to yield a lower loss. This is normally done by calculating a gradient with respect to the loss function in weight space through backpropagation [55]. Several optimizer methods exist, which are typically a variation of stochastic gradient descent [93]. Throughout this work, the Adam optimizer was used [94], together with the categorical cross-entropy loss function as implemented in the Keras framework [79]. Categorical cross-entropy is a frequently used loss function, which is defined as

$$L_{\text{crossentropy}}(\mathbf{k}^1, ..., \mathbf{k}^N; \mathbf{p}^1, ..., \mathbf{p}^N) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} k_c^i \ln p_c^i, \tag{2.3}$$

where C is the number of classes in the classification problem, and N the number of observations. $\mathbf{K}$ and $\mathbf{p}$ are the normalized ground-truth and network output vectors respectively. The number of components in both vectors equals the number of output categories of the network. Typically, only one component of the ground-truth vector $\mathbf{k}$ is one, and all the other components are set to zero, while the vector $\mathbf{p}$ contains the network output which sums to 1.

A full training cycle using the whole training dataset is usually referred to as an epoch. At the end of each epoch, the resulting model was tested on the test dataset by calculating a test loss, in order to estimate its quality on unseen data. Usually, the test loss stops decreasing or even increases again after a certain number of epochs. From this point onwards, the model is said to start over-fitting the data, i.e. learning features particular to the training set that do not improve its performance (cf. Fig. 2.10). For the ResNeXt model developed in this work, it was found that the training loss stopped decreasing significantly after approximately 15 epochs. Early stopping was used beyond that number of epochs

in order to prevent over-fitting, in accordance with common practice [95]. An analogous strategy was pursued for the sequential model.

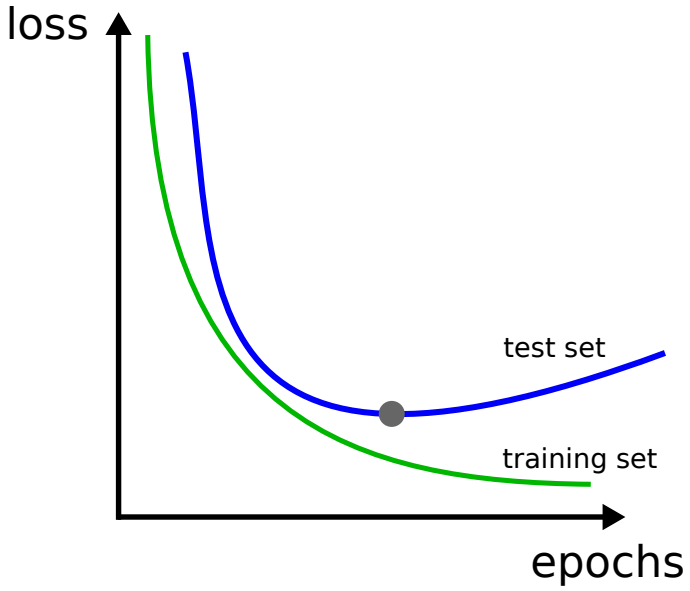


Figure 2.10: Schematic course of a typical training process of a CNN as a function of the number of training epochs. From a certain number of epochs onwards, the loss on the test set (blue line) starts to increase again, while the training loss continues to decrease. In this situation, the model is said to over-fit, i.e. learn features that do not generalize well to unseen data.

2.4.4 Network analysis

One of the strengths of CNNs is their ability to perform well on classification tasks without the explicit extraction of features of the data, which differentiates deep learning models from earlier approaches to image classification. In fact, deep learning-based algorithms have outperformed humans in a variety of defined classification tasks such as the ILSVRC, although the human performance has been found to be more robust against image degradations in some circumstances [96, 97]. However, a problematic aspect of the way CNNs produce their classification results is the difficulty to trace their predictions to particular features of the input. Hence, it is often difficult to understand and interpret network predictions, and in particular to understand the reasons why predictions fail. This issue has often been described as the "black box" property of neural networks [98]. It can be particularly problematic in the medical context, where it is often crucial not only to be able to explain and rationalise a decision, but also to assess the explanatory quality [99]. These issues have led to a whole line of research in recent years, often termed "explainable AI", which aims at better understanding and explaining the way in which machine learning algorithms reach their results [100]. In the context of image classification, a popular

method is based on calculating the gradient of the class score with respect to the input image, and was proposed by Simonyan *et al.* [101]. It allows creating co-called saliency maps which represent the importance of individual pixels in the input image for the classification decision of the network. This approach was generally taken in the present work. It should be mentioned that other approaches have been developed more recently that follow other strategies at explaining network classification decisions, including CAM [102], Grad-CAM [103], DeepLIFT [104], and LRP [105]. Comprehensive application of these different approaches to the single-cell classification problem is beyond the scope of this work. However, development of better strategies to analyse and understand neural networks, and ensure their outputs follow appropriate ethical and legal standards will be a key prerequisite to their wider use in the context of medical practice [106].

Chapter 3

Results

Evaluation of neural networks must be performed on data that has not been used in their training. While a number of methods for the analysis of the internal structure of networks can be applied, it is the performance of the network on unseen data that ultimately determines its quality as a classifier. In order to determine its practical utility, a direct comparison to human performance is necessary. In this chapter, a detailed evaluation of the trained ResNeXt and sequential networks is given in section 3.3 and 3.4. As a baseline comparison, human performance in single-cell classification as estimated using re-annotations of the dataset is described in section 3.2. Some aspects of network analysis, in particular using methods based on calculating saliency maps using the gradient of the class score, are briefly described in Sec. 3.2.

3.1 Ground truth annotation

The ground-truth annotation process performed by the first examiner on the whole scanned AOI yielded an overall database of 18,365 single-cell images of size 400 x 400 pixels, as described in detail in Sec. 2.3. For illustration of the annotation images, a part of an AOI used for ground truth annotation, as well as the resulting single-cell images are depicted in Fig. 3.1. The population of individual morphological classes according to the classification scheme used (cf. Fig. 2.4) is given in Tab. 3.1. Morphological subclasses which in the final dataset comprised fewer than 10 images were merged with neighbouring classes of the taxonomy into an overarching higher-level class. Specifically, the classes for myeloblasts with and without Auer rods were merged into a common myeloblast class, and faggot cells and promyelocytes with and without Auer rods were combined into a common promyelocyte class. Likewise, no distinction was made between polychromatic and orthochromatic erythroblasts. Merging hence resulted in a definitive structure of 15 classes for training and evaluation of the model. The resulting dataset containing the single-cell images together with ground truth annotations and re-annotations as described in Sec. 3.2 was reviewed and made publicly available by The Cancer Imaging Archive (TCIA) under Ref. [107].

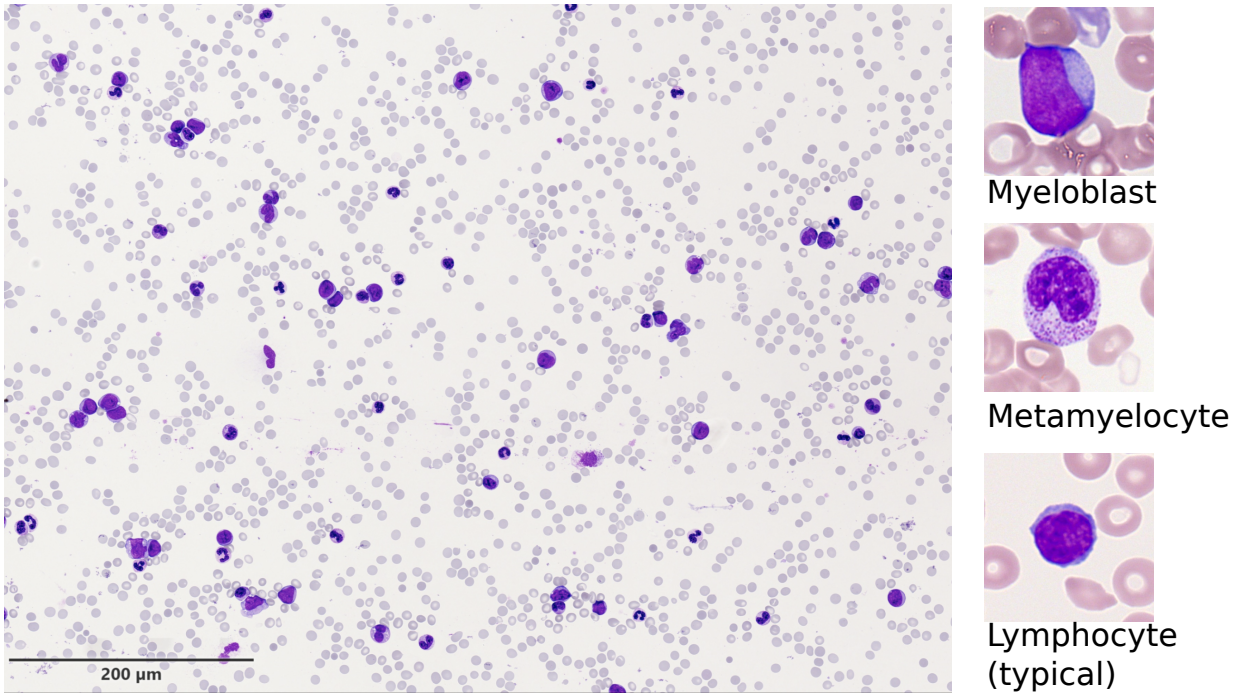


Figure 3.1: Samples of image data produced:
Left panel: Part of an area of interest (AOI) scanned from a blood smear. Leukocytes with nuclear structures can be discerned between red blood cells which do not contain nuclei. As for all samples involved in this work, the AOI scan was obtained from the monolayer of the blood smear, where cells do not aggregate or overlap significantly. Scans like the one shown here were used for the gold-standard annotation.
Figure modified from Ref. [71]
Right panels: Single-cell images extracted from the AOI scan using the ground-truth annotation given below.

3.2 Annotation quality evaluation

In order to estimate the quality of annotations, and directly compare the single-cell classification performance of a human examiner to the analogous task performed by the network, a second human examiner was asked to re-annotate a representative subsample of 1,905 single-cell images from the overall compiled dataset. Hence, the re-annotation task is precisely analogous to the classification task performed by the network. The second human examiner completed two independent re-annotation sessions, with a time separation of 11 months, during which the examiner did not have access to the sample images in order to minimise short-time memory effects. By following this strategy, two independent re-annotations of the single-cell image data subset are produced, and can be used to assess not only the agreement of the first and the second examiner, but also the consistency of a single examiner at two different points in time.
The single-cell image re-annotation task is notably different from the task of annotating

	Class	Number of images
Mature leukocytes	Neutrophil (segmented)	8,484
	Neutrophil (band)	109
	Lymphocyte (typical)	3,937
	Lymphocyte (atypical)	11
	Monocyte	1,789
	Eosinophil	424
	Basophil	79
Immature leukocytes	Myeloblast	3,259
	Myeloblast with Auer rods	9
	Promyelocyte	67
	Promyelocyte with Auer rods	1
	Faggot cell	2
	Promyelocyte (bilobed)	18
	Myelocyte	42
	Metamyelocyte	15
	Monoblast	26
	Erythroblast	78
	Smudge cell	15
	Total	18,365

Table 3.1: Full class-wise statistics as annotated by the first examiner. To ensure a sufficient number of images for training and testing, subclasses containing less than 10 cells were merged as described in the main text (cf. Table 1 of the main text).
Table reproduced from Ref. [71].

individual cells on the whole scanned AOI, which is performed by the first examiner. While in the re-annotation task, individual cell images are annotated out of context, the initial AOI-based annotation task enables simultaneous assessment and comparison of all cells on the AOI scanned. The annotations obtained from a whole-AOI scan include global information of the smear, e.g. a comparison between different cell types present, and are therefore expected to be more accurate. For this reason it seems appropriate to regard the first annotation performed on the whole scan as a gold-standard, and use it as ground truth when training and evaluating the networks as presented in this chapter. The second examiner had the possibility to mark single-cell images for which a definite morphological type could not be determined as "unclear" in both re-annotation sessions.

3.2.1 Inter-rater agreement

The results of the first and second re-annotation round are presented in the left column of Fig. 3.2, as compared to the ground truth provided by the first human examiner. Both re-annotations show a similar deviation pattern from the ground truth label, with signif-

icant deviations mainly focussed within the consecutive steps of myelopoiesis. Normally, deviations within these cell classes are considered tolerable and unproblematic, as no exact morphological criterion exists that exactly delineates the consecutive substeps of myeloid development. In fact, previous work by Krappe *et al.* suggested treating respective classes as equivalent [108]. Further deviations of the re-annotations from the ground-truth annotation concern myeloblasts confused with lymphocytes or monocytes, which seems morphologically plausible.

A subset of 63 single-cell images (3.3% of the re-annotation dataset), and 208 single-cell images (10.9% of the re-annotation dataset) could not be assigned a unique morphological label by the second human cytologist during the first re-annotation and the second re-annotation 11 months later respectively. The difference in the number of these unlabelled single-cell images suggests that examiner confidence in classifying single-cell images may vary over time. The right column of Fig. 3.2 shows the distribution of ground-truth labels of single-cell images marked as "unclear" by the second examiner. While the number of cells labelled "unclear" was substantially larger in the second re-annotation, we note that similar cell types are affected in both re-annotations, namely myeloblasts, typical lymphocytes, monocytes and segmented neutrophils, again reflecting cell types that are prone to be confounded according to the confusion matrix. For a confident differentiation of those morphological cell types, the contextual information on the whole AOI scan may hence be particularly useful.

Within the group of images for which the second examiner did assign a unique morphological class, re-annotations showed excellent agreement with the ground-truth label. A common metric for quantification of inter-rater agreement is Cohen's kappa, which is defined as [109]

$$\kappa = \frac{p_0 - p_e}{1 - p_e}, \tag{3.1}$$

where p_0 represents the observed agreement between raters, and p_e the overall proportion of chance-expected agreement. Note that $\kappa = 0$ for totally coincidental agreement between raters, and $\kappa = 1$ for perfect agreement. In practice, p_0 is calculated from the elements of the confusion matrix c_{ij} as the relative frequency of samples for which both raters agree out of a total of N samples as

$$p_0 = \frac{\sum_i c_{ii}}{N}, \tag{3.2}$$

while p_e is determined as

$$p_e = \frac{1}{N^2} \cdot \sum_i c_{i\cdot} \cdot c_{\cdot i}, \tag{3.3}$$

where the $c_{i\cdot} = \sum_j c_{ij}$ and $c_{\cdot i} = \sum_j c_{ji}$ are the marginal populations of the confusion matrix. Specifically, the value of Cohen's kappa for the re-annotation single-cell image labels (excluding "unclear" cases) compared to the gold-standard was $\kappa = 0.84$ for the first, and $\kappa = 0.87$ for the second re-annotation, indicating excellent agreement in both re-annotations according to common interpretations of κ [109].

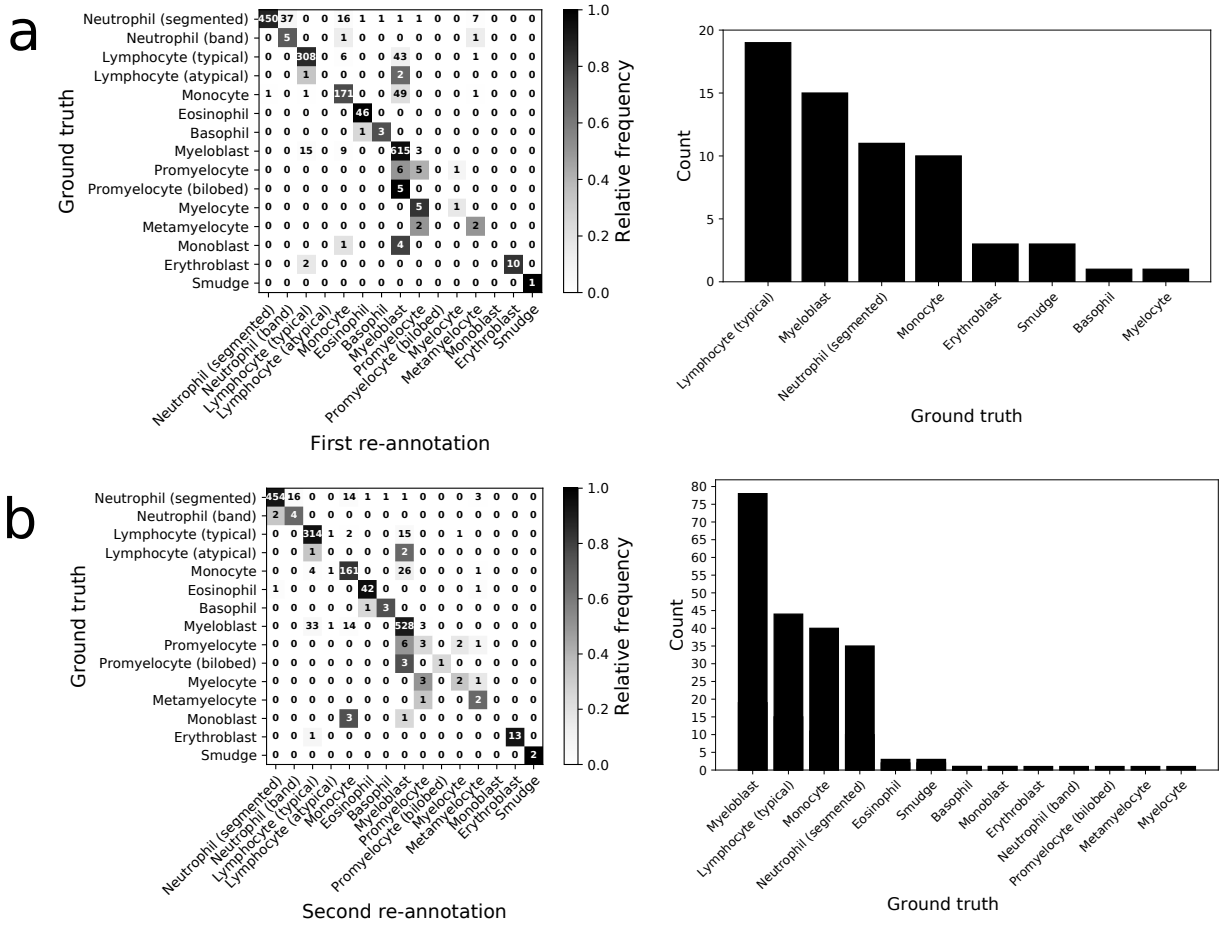


Figure 3.2: Re-annotations of single-cell images performed by a second human examiner compared to the ground truth provided by the first examiner with access to the whole AOI scan.
(a) Confusion matrix between ground-truth label and first re-annotation label for uniquely classified images (left), and ground-truth labels of the 63 images marked as "unclear" during the first re-annotation round (right). Excellent agreement is observed with a value of COHEN's kappa of $\kappa = 0.84$.
(b) Confusion matrix (left) and distribution of ground-truth labels of the 208 images marked "unclear" (right) in the second re-annotation round, performed with a time distance of 11 months from the first re-annotation shown in (a). Again, excellent agreement is obtained with a value of COHEN's kappa of $\kappa = 0.87$. In both re-annotations, note similar deviations from the ground truth are observed, for example in classifying atypical lymphocytes or promyelocytes as myeloblasts.
Figure reproduced from Ref. [71]

3.2.2 Intra-rater agreement

Rather than comparing the results of the single-cell image re-annotations provided by the second examiner to the ground-truth annotation of the first examiner as was done in the previous section, both re-annotations can be directly compared in order to assess the self-agreement of a single human examiner performing the analogous task of the neural network. The two respective re-annotation sessions were carried out by the same examiner and separated by a time gap of 11 months. This enables a direct comparison of both annotations to estimate the intra-examiner variability of the single-cell image annotation process. Intra-examiner variability may be taken as to estimate the day-to-day variability of classification performance, which is known to potentially be a considerable source of imprecision [110]

A comparison of both singe-cell annotations of the second examiner is given using a confusion matrix in Fig. 3.3. With a COHEN's $\kappa = 0.77$, results also for this measure lie in the domain of excellent agreement [109], showing good consistency of second annotator performance over time.

3.3 ResNeXt model evaluation

This section presents the ResNeXt model as trained and evaluated using the entire dataset, and distributing it randomly to 5 folds in a stratified way, as described in Sec. 2.4.3. The model is evaluated for two tasks, namely classification of images according to the morphological scheme described in Sec. 2.3, and binary classification.

3.3.1 Classification performance

Performance of the ResNeXt network is evaluated by passing single-cell images through it, and comparing the output prediction with the ground-truth labels assigned by the ground truth examiner (cf. Sec. 2.3). In its final layer, the network outputs a vector of normalised activation

$$\mathbf{P} = (P_1, ..., P_i, ..., P_{15}), \tag{3.4}$$

which can be interpreted as predicted probabilities for the input to belong to a respective class. Here, the component P_i correspond to the predicted probability for the given image to belong to class i out of the 15 overall classes. The network's classification prediction is the class m with the highest predicted probability P_m.

Class-wise predictions of the ResNeXt are shown in the confusion matrix of Fig. 3.4. Additionally, the class-wise performance of the model can be described using common metrics for evaluation of diagnostic tests, namely precision, sensitivity and specificity, which are

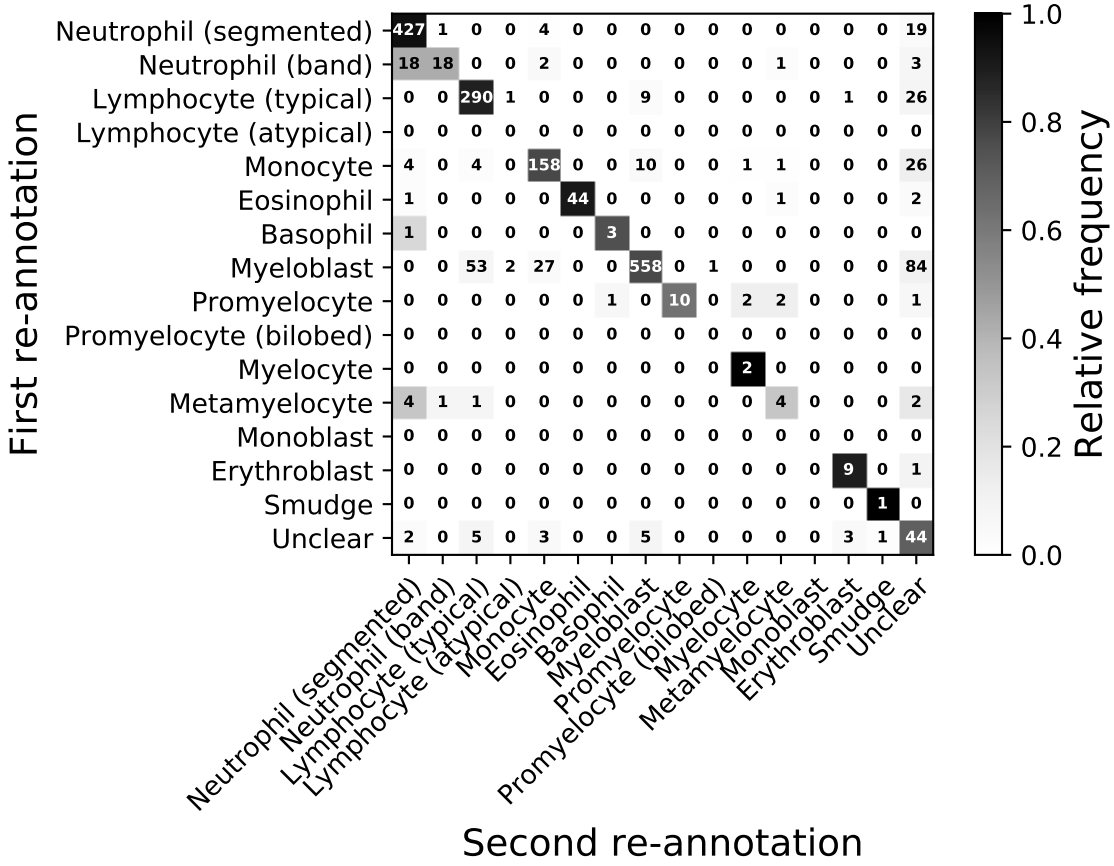


Figure 3.3: Excellent agreement between two annotations (82.3% overall, $\kappa = 0.77$) of a subsample 1,905 single-cell images. Both re-annotations were produced by the same human examiner with an intervening time of 11 months.
Figure reproduced from Ref. [71].

commonly defined as follows:

$$\text{sensitivity} = \frac{\text{true positive}}{\text{positive}} \tag{3.5}$$

$$\text{specificity} = \frac{\text{true negative}}{\text{negative}} \tag{3.6}$$

$$\text{precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}}. \tag{3.7}$$

In this context, “true positive” and “true negative” are the number of images correctly ascribed or not ascribed to a given class by the network, respectively, “positive” and “negative” are the overall number of images belonging or not belonging to a certain class, and “false positive” is the number of cell images wrongly ascribed to that class. Values of

precision and sensitivity for all cell classes obtained by 5-fold cross-validation are given in Tab. 3.2.

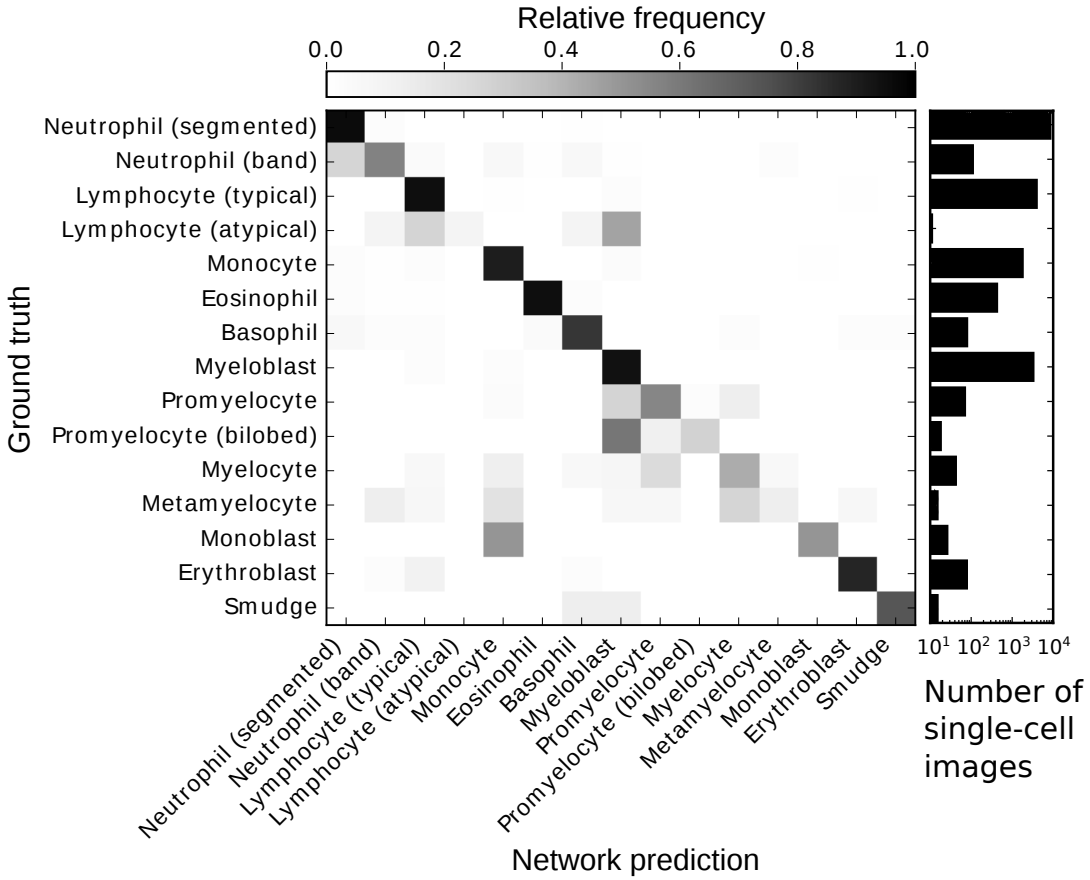


Figure 3.4: Confusion matrix between network prediction and ground-truth label of the human examiner obtained by 5-fold cross-validation. To the right of the matrix, the number of single-cell images in the overall dataset is indicated on a logarithmic scale. For key cell classes such as myeloblasts, the ResNeXt network shows very good performance, with deviations from ground-truth similar to the human second examiner discussed in Sec. 3.2. In Tab. 3.2 the class-wise performance is given using precision and sensitivity. Individual results for all 5 folds are given in Appendix B.
Figure adapted from Ref. [71].

Network predictions agree very well with gold-standard annotations for the most common physiological cell types, e.g. segmented neutrophils, typical lymphocytes, monocytes, and eosinophils, achieving values above 90% in both precision and sensitivity in these classes (cf. Tab. 3.2). Myeloblasts, whose presence in the peripheral blood is common in myeloid leukemias [20], are also recognised with a very high precision and sensitivity of 94% (cf. Tab. 3.2).
Rare classes are more challenging for the network to classify correctly, in particular the

	Class	Precision	Sensitivity	Number of images
Mature leukocytes	Neutrophil (segmented)	0.99 ± 0.00	0.96 ± 0.01	8,484
	Neutrophil (band)	0.25 ± 0.03	0.59 ± 0.16	109
	Lymphocyte (typical)	0.96 ± 0.01	0.95 ± 0.02	3,937
	Lymphocyte (atypical)	0.20 ± 0.4	0.07 ± 0.13	11
	Monocyte	0.90 ± 0.04	0.90 ± 0.05	1,789
	Eosinophil	0.95 ± 0.04	0.95 ± 0.01	424
	Basophil	0.48 ± 0.16	0.82 ± 0.07	79
Immature leukocytes	Myeloblast	0.94 ± 0.01	0.94 ± 0.02	3,268
	Promyelocyte	0.63 ± 0.16	0.54 ± 0.20	70
	Promyelocyte (bilobed)	0.45 ± 0.32	0.41 ± 0.37	18
	Myelocyte	0.46 ± 0.19	0.43 ± 0.07	42
	Metamyelocyte	0.07 ± 0.13	0.13 ± 0.27	15
	Monoblast	0.52 ± 0.30	0.58 ± 0.26	26
	Erythroblast	0.75 ± 0.20	0.87 ± 0.09	78
	Smudge cell	0.53 ± 0.28	0.77 ± 0.20	15
	Total			18,365

Table 3.2: Class-wise precision and sensitivity of the network, determined by 5-fold cross-validation. The model attains levels of precision and sensitivity above 0.9 for morphological classes containing more than 400 images, such as segmented neutrophils, typical lymphocytes and myeloblasts. Larger deviations across folds occur for classes with small sample number, e.g. metamyelocytes and promyelocytes.
Table reproduced from Ref. [71].

intermediate stages of granulopoiesis and erythropoiesis, and basophils, for which our test and training dataset contains less than 100 images. Note that these mixups within granulopoiesis were also observed in the performance of the human re-annotator described in Sec. 3.2. To some extent, this effect is due to the lacking precise morphological delineation between the different maturation stages. Therefore, these mixups have been considered as tolerable in the literature [111].

Due to the intrinsically unbalanced number of cells present in the scanned smears for different cell types, the number of test and training images varies by up to two orders of magnitude for different classes (cf. Tab. 3.2). As might be expected, the number of single-cell images available in a particular class of the dataset correlates with network performance. E.g., as inspection of Tab. 3.2 shows, the model attains values above 0.9 in classes for which more than 400 single-cell images are present in the dataset.

As a further consequence of the high intrinsic class imbalance in the single-cell image dataset, calculation of an overall accuracy score for our model is problematic, as it would be biased towards the classes with a high number of samples [55], and is therefore not evaluated here.

Values in Tab. 3.2 as well as the entries in the confusion matrix of Fig. 3.4 were obtained by 5-fold cross-validation as described in Sec. 2.4.3, in order to make results less dependent on the random noise introduced by the allocation of cells into individual folds. A description of the individual results of the 5 different models trained for cross-validation is given in

Appendix B.

3.3.2 Binary decision performance

From the classification of single-cell images into distinct morphological categories, coarser classifications can be derived by summing up the activations for several categories. Specifically, by taking the sum

$$P_{\text{blast}} = P_{\text{myeloblast}} + P_{\text{monoblast}}, \tag{3.8}$$

the overall probability of a cell to possess blast character can be calculated from the network output. The question whether myeloblasts or monoblasts are present on a blood smear possesses key clinical importance, as these so-called blast equivalents are generally required to be present in the peripheral blood for a diagnosis of AML [20]. Given the network output of P_{blast}, a threshold probability t can then be chosen such that the binary prediction of the network is given by

$$\hat{y} = P_{\text{blast}} \geq t. \tag{3.9}$$

The receiver operating characteristic (ROC) curve is the result of sweeping t between 0 and 1, and is shown in the upper panels of Fig. 3.5.
Averaging across the 5 folds trained, the area under the ROC curve is obtained as AUC-ROC $= 0.992 \pm 0.001$. Hence, ResNeXt provides a test of the blast character of a given single-cell image that fulfills the criteria of outstanding tests using the usual criteria of test assessment [112, 113]. To relate the network's performance with the human examiner, sensitivity and specificity of single-cell re-annotation can be assessed by considering if the re-annotator classified an image as either myeloblast or monoblast. Human performance exhibits a sensitivity of 95.7% and 90.7%, and a specificity of 91.1% and 95.2% for the first and second re-annotation respectively (cf. upper panels of Fig. 3.5). Results of both independent, human re-annotations lie close to but somewhat below the network ROC curve. This indicates that the network achieves a comparable and slightly superior performance compared to the human examiner in deciding if a given single-cell image contains a blast-like cell or not.
In analogy to the blast vs. non-blast decision, another clinically important binary decision on individual leukocytes is whether a given cell belongs to one of the typical cell types present in peripheral blood under normal circumstances, or to atypical cell types that occur in pathological situations. In this context, atypical cells are myeloblasts, monoblasts, myelocytes, metamyelocytes, promyelocytes, erythroblasts and atypical lymphocytes. In this test, the overall probability P_{atypical} for a given cell to be classified as atypical is obtained by summing up the output probabilities of all individual atypical cell classes. Again, the ROC is determined by sweeping through threshold values t for the atypicality test as given Eq. 3.9 for the blast test.
The ResNeXt network yields an ROC-AUC of 0.991 ± 0.002 when testing for atypicality of a cell, which again indicates outstanding performance [112, 113]. In comparison, human re-annotation attains a sensitivity of 95.9% and 91.7% and a specificity of 91.0% and 95.3% for the first and second re-annotation respectively (lower panels of Fig. 3.5). It can be observed that the sensitivity–specificity point of the human examiner lies slightly below the ROC curve produced by the network. Hence performance by the ResNeXt model is

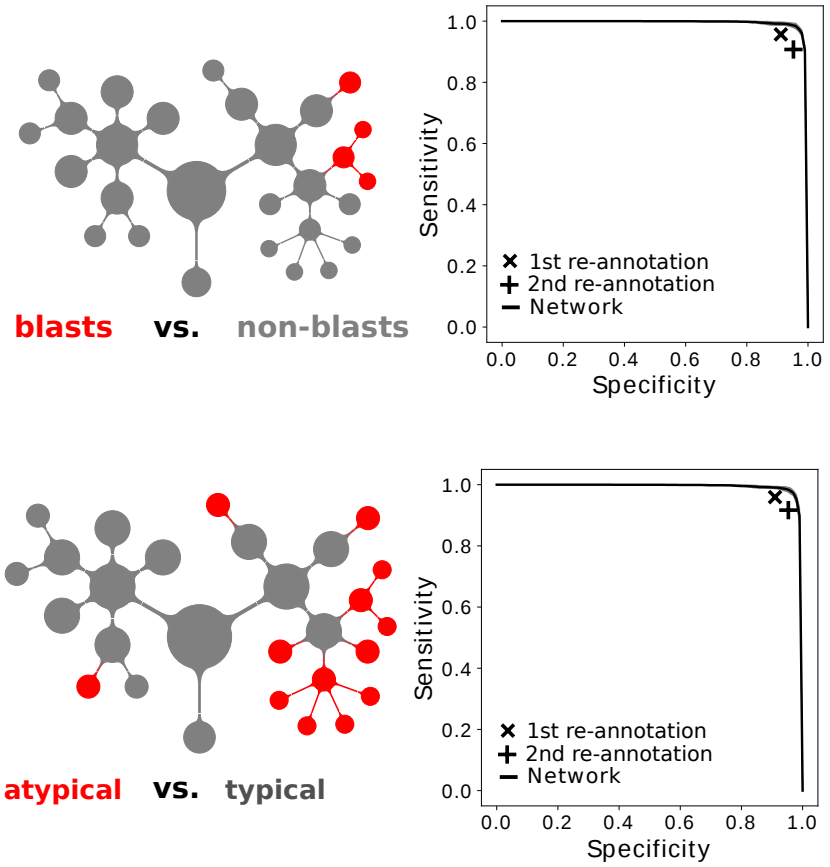


Figure 3.5: Comparison of ResNeXt model and human performance on binary tasks. The morphological classes for which the respective binary parameter is positive are highlighted red in the taxonomy schematic in the left column (cf. Fig. 2.4).
Upper panels: Model ROC for the binary test for blast character of a given single-cell image. The network performs very well with an area under the curve (AUC) of 0.992 ± 0.001, measured by averaging across five folds. Indicated by indicated by 'x' and '+' are performances of the human second examiner during two independent re-annotations of single-cell images at different times, which are slightly outperformed by the network.
Lower panels: ROC of the ResNeXt network in the binary task of recognising atypical cells. Also in this task, the network performs very well (AUC: 0.991 ± 0.002) and slightly outperforms the human second examiner.
Figure adapted from Ref. [71].

close to and slightly better than the human examiner's performance in both re-annotation rounds on the single-cell annotation task relative to the ground truth on our test subset containing 1,905 images. This observation holds true for both the blast character and atypicality tests, as can be seen from Fig. 3.5.

3.3.3 Alternative training regime

For the training process of the network in Secs. 3.3.1 and 3.3.2, all annotated single-cell images from the dataset were pooled prior to performing the class-wise, stratified split described in Sec. 2.4.3, which assigned approximately 20% of the images in a specific class for testing and the remaining 80% for training in each fold. All images used as in the test set were excluded from the training process and not seen by the network before testing. This method of dividing train and test samples allows distributing the available image pool evenly between test and training sets. However, different single-cell images from the same blood smear can be assigned to the training and test classes in the same split. Hence *a priori*, one might worry if this method could introduce a bias into the test set due to correlations between images of different single cells stemming from the blood smear of the same patient, e.g. by stain or focus effects shared amongst images that come from the same slide.

In order to test the importance of possible correlations of this kind, the model was re-trained, this time splitting the training and the test sets according to patient-of-origin rather than cell type. Specifically, all cells from 10 patients with AML diagnoses and 10 patients without pathological peripheral blood smears were set aside for testing. No data from these patients was used in the training process of the alternative network model. This procedure restricts possibilities of an even split within the individual cell classes, which is particularly problematic for classes for which only few sample images exits. If a patient who contributes many samples of a rare cell type gets selected into the test or training set, this implies a substantial decrease in the number of available images in the other set, making it difficult to train or test the network for that particular class reliably.

Despite this complication, the network trained by the patient-wise train-test split performs well for all major classes, as can be seen from Fig. 3.6, which shows evaluations of the alternative model analogous to Figs. 3.4 and 3.5. The AUC values of the case-wise model for the binary tasks of distinguishing blasts from non-blasts and typical from atypical cells are 0.992 and 0.986 respectively. Good agreement between the class-wise and the patient-wise data splitting regime indicates that correlations between different leukocytes on a slide are not significant enough to determine training performance. Leukocytes so not appear to carry “hidden name tags” that indicate which slide they are imaged from. However, class-wise splitting in test and training set allows making maximal use of the dataset, as it does not introduce conditions on single-cell distribution.

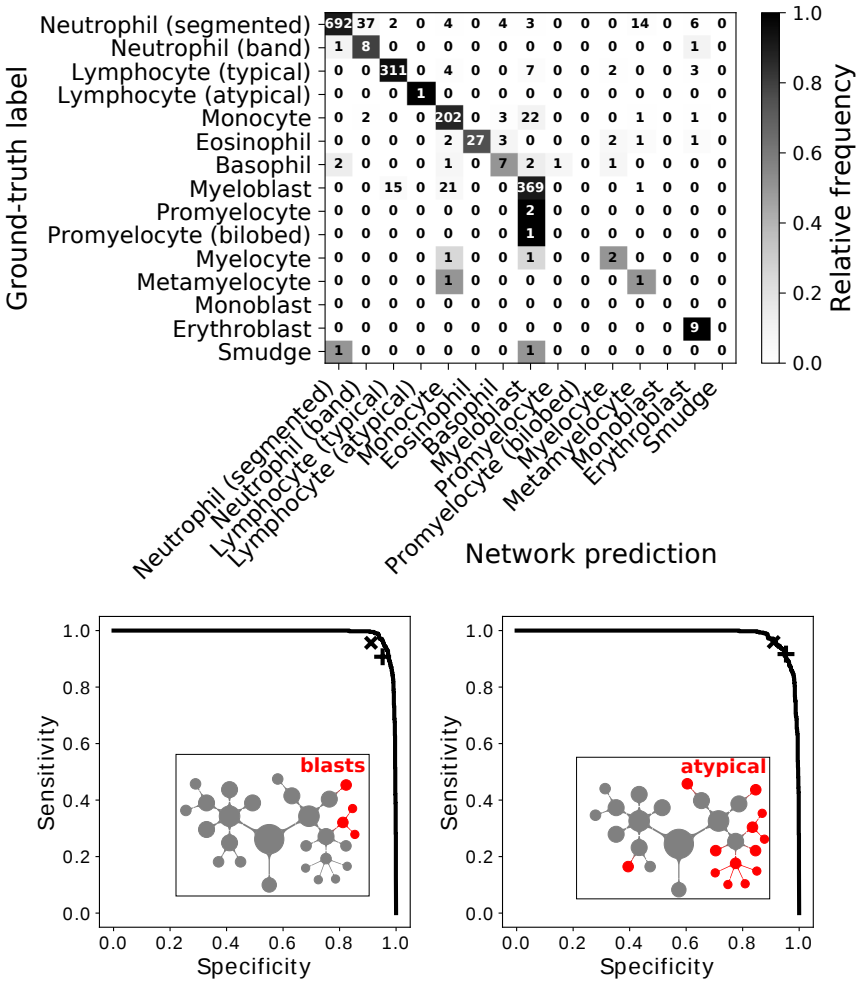


Figure 3.6: Predictions of a ResNeXt model trained by splitting between training and test set according to patient-of-origin rather than morphological class.
Upper panel: Confusion matrix of the ResNeXt-model trained on a patient-wise split of test and training data. Classification results are similar to the model trained based on a stratified class-based split (cf. Figs. 3.3.1). However, for classes containing only a small number of single-cell images, deviations occur due to the difficulty of separating them into test and training set on a case-wise basis. Notably, all monoblast cells in the dataset stem from a single case, so that no case-wise split into training and test sample is possible.
Lower panels: Left: ROCs of the case-wise model when testing for blasts. At this task, the model achieves an AUC of 0.992, just as the model presented in Sec. 3.3 that was trained based on a class-based split. Right: ROC of the ResNeXt model trained on a patient-based split when distinguishing typical from atypical cells. The model achieves an AUC of 0.986 for this task. Insets schematically depict the classification taxonomy, with the classes contributing to the respective binary decisions coloured red.
Figure adapted from Ref. [71].

3.4 Sequential model evaluation

Neural networks typically possess a large number of hyperparameters, which have to be adjusted according to the problem at hand. Finding their optimal values is generally a difficult problem, which in many cases can only be done in an iterative way or by trial and error, as systematic optimization, e.g. by grid search, is often too computationally expensive. In order to show the robustness of the results obtained using the ResNeXt model, a second network is trained here with a different architecture. Namely, a sequential model is used, which is inspired by the VGG model [85]. Details on architecture and hyperparameter choice in this second network are described in Sec. 2.4.2.
The sequential model is trained to perform precisely the same image classification task using exactly the same split of the data into 5 folds as for the ResNeXt model (cf. Sec. 3.3). Hence, a direct comparison of the results of both networks is possible, with differences in prediction performance entirely due to differences in network architecture. Note that the sequential model possesses only 433,224 trainable parameters, which compares to 23,046,800 trainable parameters for the ResNeXt model and hence differs by a factor of more than 50. In the sequential network, a 16-class classification was originally performed, with a further differentiation of orthochromatic and polychromatic erythroblasts. For consistency with the results of ResNeXt model, these two classes are lumped into one common erythroblast class.

3.4.1 Classification performance

Performance of the sequential model in the classification task compared to the ground-truth annotation is shown in Fig. 3.7. The confusion matrix was obtained using 5-fold cross-validation also for this model. The general deviation pattern is similar to what was observed in the ResNeXt model (cf. Fig. 3.4). In particular, similar mixing between the individual stages of granulopoiesis is apparent, which as discussed for the ResNeXt model in Sec. 3.3 is intrinsic to the morphological classification process and also present in the results of human re-annotation (cf. Fig. 3.2). The class-wise values of precision and sensitivity obtained by the sequential model are listed in Tab. 3.3.

3.4.2 Binary decision performance

For the two binary questions considered in this work, namely discerning blasts from non-blasts and typical from atypical cell morphologies, performance of the sequential model is depicted in Fig. 3.8. Also the sequential model performs very well, with an ROC-AUC of 0.983 ± 0.003 and 0.990 ± 0.001 for recognition of blast cells and atypical cells respectively. Again, performance of two re-annotations by the second human examiner is close to, but somewhat below the ROC, indicating that also the sequential model performs at least on a par with the human performance. Hence, as far as blast recognition and the recognition of atypical cell types is concerned, the sequential model is only very slightly inferior to ResNeXt. This fact is remarkable given the much higher model size of ResNext, and

	Class	Precision	Sensitivity	Number of images
Mature leukocytes	Neutrophil (segmented)	0.99 ± 0.04	0.94 ± 0.01	8,484
	Neutrophil (band)	0.16 ± 0.06	0.60 ± 0.17	109
	Lymphocyte (typical)	0.97 ± 0.02	0.93 ± 0.02	3,937
	Lymphocyte (atypical)	0.11 ± 0.11	0.11 ± 0.27	11
	Monocyte	0.88 ± 0.02	0.90 ± 0.03	1,789
	Eosinophil	0.91 ± 0.08	0.94 ± 0.02	424
	Basophil	0.45 ± 0.14	0.75 ± 0.06	79
Immature leukocytes	Myeloblast	0.94 ± 0.03	0.92 ± 0.03	3,268
	Promyelocyte	0.49 ± 0.12	0.50 ± 0.17	70
	Promyelocyte (bilobed)	0.21 ± 0.07	0.69 ± 0.30	18
	Myelocyte	0.50 ± 0.19	0.62 ± 0.20	42
	Metamyelocyte	0.14 ± 0.13	0.40 ± 0.44	15
	Monoblast	0.37 ± 0.14	0.67 ± 0.33	26
	Erythroblast	0.59 ± 0.20	0.88 ± 0.09	78
	Smudge cell	0.53 ± 0.33	0.67 ± 0.21	15
	Total			18,365

Table 3.3: Class-wise precision and sensitivity of the sequential network, determined using 5-fold cross-validation.
Also the sequential model achieves precision and sensitivity above 90% for most of the key diagnostic cell classes, including segmented neutrophils, typical lymphocytes and myeloblasts. As in the case of the the ResNeXt model, deviations across folds occur for classes with small sample number, e.g. metamyelocytes and promyelocytes.
Table reproduced from Ref. [71].

suggests that, as suspected in Sec. 2.4.2, model performance is limited by the size of the dataset and internal annotation consistency, rather than model size.

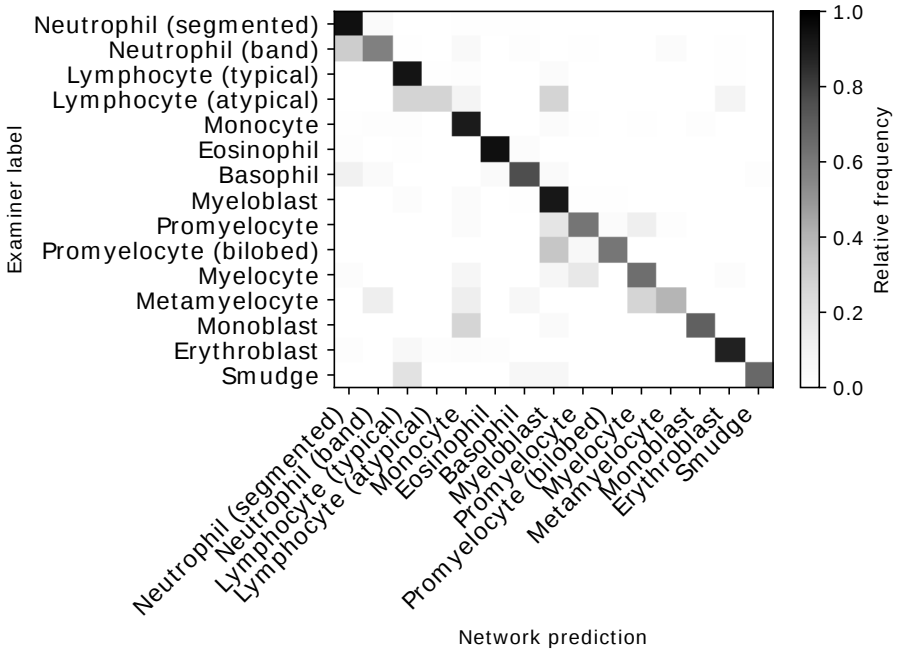


Figure 3.7: Classification performance of a sequential model trained and tested on the same dataset as the ResNeXt model (cf. Sec. 3.3 and Fig. 3.3). The confusion matrix for the morphology classification task was obtained by 5-fold cross-validation.
Figure modified from Ref. [71].

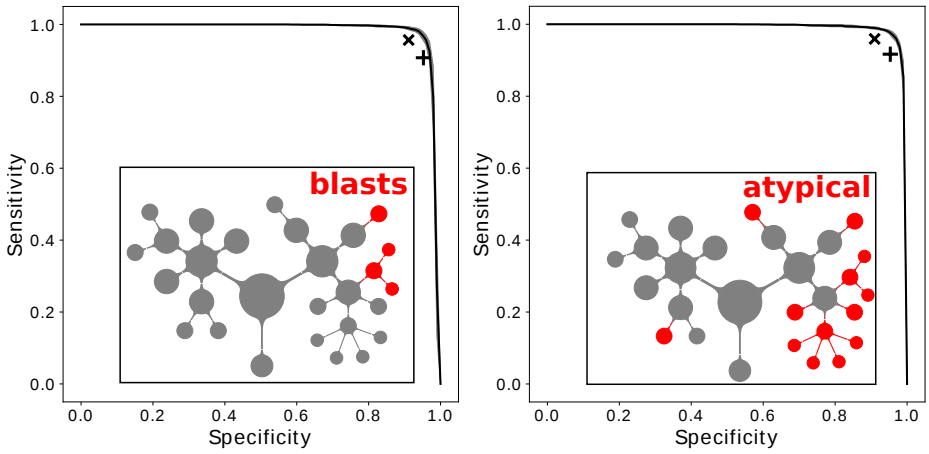


Figure 3.8: Performance of the sequential model in binary decisions. Insets show schematic classification taxonomy, with positively contributing classes labelled red. Left panel: ROC of the network when distinguishing blasts from non-blasts. This binary classification is only slightly inferior to the performance of the ResNeXt network, achieving an AUC of 0.983 ± 0.003. Left panel: ROC of the sequential network distinguishing typical from atypical cells. Also for this task, performance is good with an AUC of 0.990 ± 0.001.
Figure modified from Ref. [71].

3.5 Model analysis

To determine if the network focusses on relevant parts of the single-cell images, saliency maps were calculated for both models following the gradinent-based procedure outlined by Simonyan *et al.* [101]. With the help of that method, it is possible to track and visualise how important each pixel is individually for the output of the respective network, i.e. the classification decision. Saliency maps for several test-set images are shown in Fig. 3.9 for the ResNeXt model and in Fig. 3.10 for the sequential model, which were both trained using precisely the same dataset.
From these maps, it can be inferred that pixels within the leukocyte classified are of key importance for the network's classification decision. This observation suggests that the network has learned to focus on the areas of the single-cell image that are known to be relevant for differentiating leukocytes, e.g. nuclear shape and the staining behaviour of the cytoplasm, which emulates the known criteria for morphological differentiation [9].
No obvious correlation could be discerned between the saliency maps and the result of the classifications as compared to the ground truth, suggesting that both correct and incorrect classifications were obtained by the network by focussing on the single-cell image region which contains the leukocyte. Hence, errors do not solely occur by the network focussing on a background aspect of the single-cell image patch. This overall behaviour appears for both networks, which is compatible with their observed similar performance on the respective test sets. In the ResNeXt case, cell boundaries appear to show a slightly sharper delineation in the saliency maps, which may indicate that this model possesses a better edge detection capacity.
Observations of the pixel-wise classification behaviour of images as depicted using saliency maps help illustrating the network's behavior, and may can be regarded as a "sanity check" on the trained network. An illustration that the model decisions rest on similar parts of the input as human decisions would may increase trust in the model output. However, these visualizations do not offer a proof of model behaviour, and cannot be used to predict the general performance of the network on unseen data.
As discussed in more detail in Sec. 2.4.4, a variety of alternative methods have been proposed in recent years aimed at allowing an "anatomy of neural networks" and analyse the way in which network output is produced from the inputCAM [102, 103, 104, 105]. These methods allow gaining a better understanding of the inner structure of the model. Additionally, methods have been developed that aim at making the comparison between saliency maps more quantitative and rigorous [114]. Systematic application of these different techniques is beyond the scope of this work. However, this direction of research continues to be an active and important complement to the development of neural networks, with the potential of building trust in the decision making process of networks by offering a glimpse at their inner workings, and suggesting improvements to existing schemes. In particular, analysis of the way in which network reach arrive at their output predictions can help to make sure that these are not generated by biases or random associations in the training data. Finally, this set of methods may also be able to show which parts of the input data enable the network to reach the high level of performance it exhibits in the image classifi-

cation task. Knowledge of these "criteria" applied by the neural network may enable the human examiner to learn about so-far ignored aspects of the dataset, and generate new hypotheses on it.

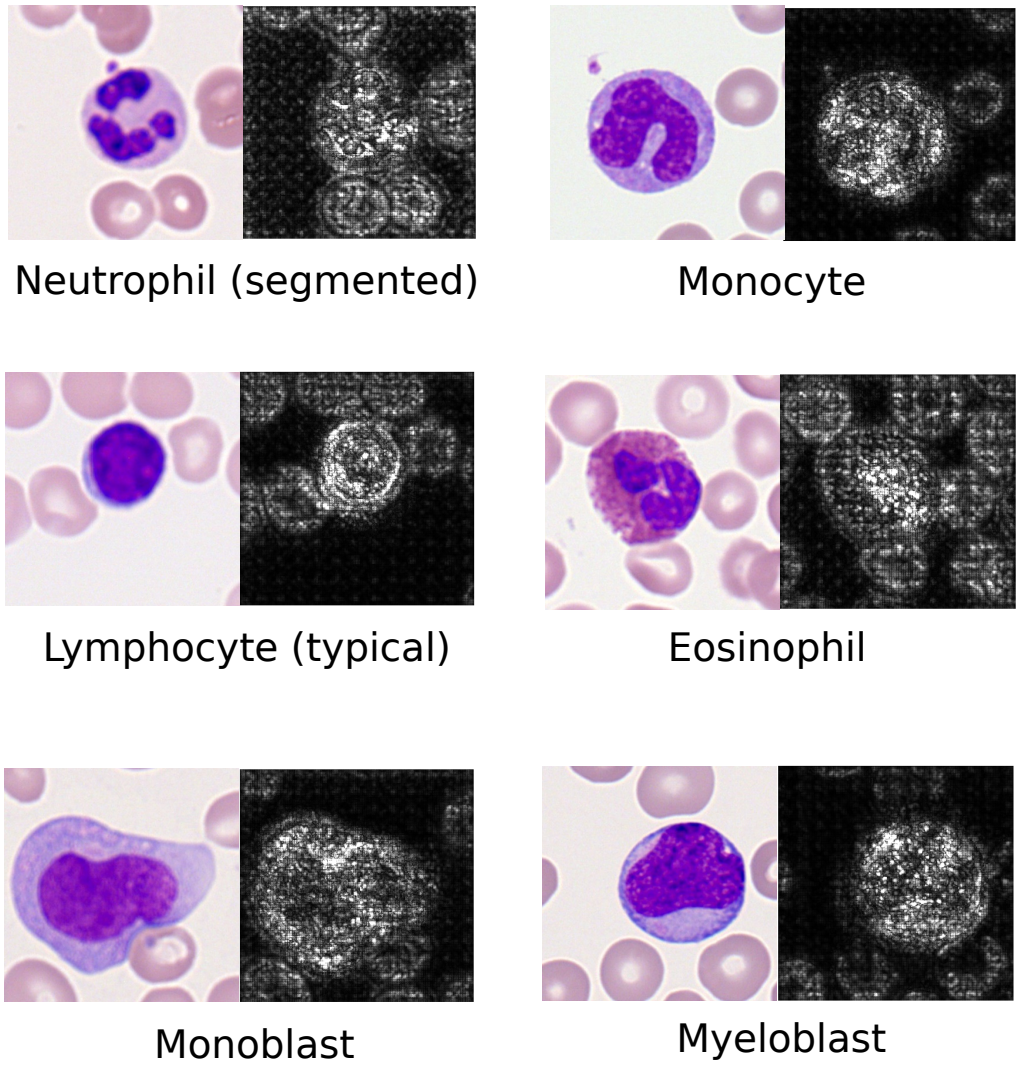


Figure 3.9: Saliency maps illustrate the gradient of a pixel with respect to the ResNeXt model's loss function. Brighter pixels have a higher influence on the network's classification decision. Maps suggest that the network learns to focus on the leukocyte and map out its internal structures, such as nuclear shape and cytoplasmatic content, while giving less weight to background content.
Figure modified from Ref. [71].

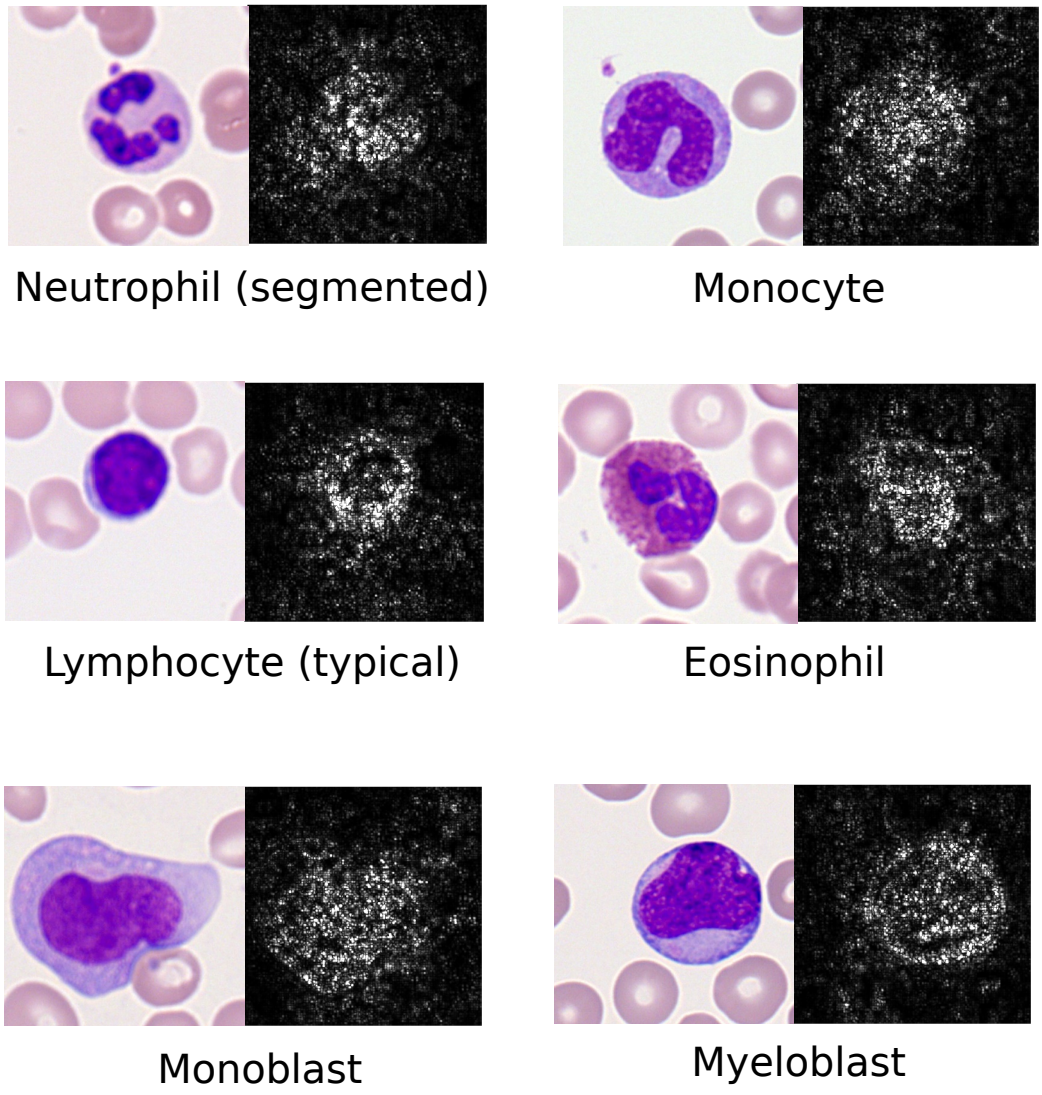


Figure 3.10: Saliency maps of the sequential model trained and evaluated on the same images as the ones shown in Fig. 3.9. Also for the sequential model presented in this work, the highest-weight regions of the saliency maps lie within the leukocyte, indicating again that the model focusses on relevant regions also in this network. Saliency maps for the sequential model seem to map out the cell boundaries less sharply, which might indicate a higher edge detection capacity in the much more complex ResNeXt model.

Chapter 4

Discussion and Outlook

4.1 Implications of the present work

In the present thesis, an annotated dataset of leukocyte cytomorphologies was compiled. To the author's knowledge, it is presently the largest publicly available dataset on leukocyte morphologies in leukemias. Images contained in the dataset were annotated up to three times, which allowed for a careful analysis of examiner performance and annotation quality. The dataset has been made publicly available in order to serve as a reference for future endeavours in the field of leukocyte image analysis.

In the context of the present work, the dataset proved large enough to allow application of neural networks, a highly data-driven machine learning method. The trained networks are able to classify images of single leukocytes into a standard morphological classification scheme. Specifically, the two CNNs presented in this thesis exhibit very good performance at identifying the most important morphological white blood cell types present in peripheral blood without morphologic signs of malignancy. Both network structures also perform very well at identifying pathological cell types which are key in the diagnosis of acute myeloid leukemia. For the most common physiological leukocyte classes as well as for myeloblasts, the networks attain precision and sensitivity above 90% when probed on test datasets. Hence, these cell types can be identified with a very high accuracy, outperforming other classifiers developed so far in the literature by a significant margin [115, 111]. Classification predictions of the networks have been used to answer clinically relevant binary questions, in particular if a given cell possesses blast character or if it is atypical and normally absent under physiological conditions. It was found that in these two binary decision tasks, the networks both reach the performance of human examiners classifying individual leukocytes. Analysis of the trained networks using saliency maps hints that the networks have learned to reach their classification decision by focussing on parts of the image which are also being examined by human cytologists. While this observation does not in it self validate the model, it might be taken as additional, positive check of its properties.

Given this level of performance and the fact that the method is easily scalable and fast, the algorithm can be used to quickly screen thousands of cells on a blood smear scan, helping

cytologists to find suspicious cells more readily. Rapid, automated pre-screening might be particularly helpful in situations where the number of malignant cells is expected to be small, such as in the early stages of hematological diseases, or at the beginning of relapse. For fully automatic screening of all leukocytes on a blood smear, the single-cell classification method developed here can be combined with a segmentation tool that selects single cells. For this task, a large number of algorithms are available, which could be combined with the cell classifier [116, 117, 118, 119]. As developed in this work, the network hence has the potential to act as a rapid pre-screening and quantitatively informed decision tool for cytological examiners which is based on the expertise of experienced examiners and can speed up routine diagnostics.

4.2 Challenges for deep learning models in leukemia diagnostics

The model developed in this work has shown very good performance on the single-cell image classification task, showing human-level results in recognising blast-like and atypical cells and outperforming other approaches to this problem used so far. Nevertheless, further tests and evaluations of the model are necessary.

First, as might be expected for a data-driven classification method, a correlation was found between the number of images available for a specific class in the training data set and the performance of the network on that class. The single-cell image dataset presented in this work containing over 18,000 single-cell images from 200 patients is considerably larger both in terms of images and of patients included than other datasets available so far. However, compared to databases like ImageNet, the dataset is still relatively small. This may also be inferred from the fact, that the much simpler sequential model performs almost on the same level as the much more complex ResNeXt model, which might be able to play out its relative strengths only when used on a large dataset. Hence, from experience with the general behaviour of neural networks, one can anticipate that further enlarging the dataset will further improve the networks' classification performances, in particular for cell types which are rare in the present dataset.

Secondly, as neural networks learn from the input data in an end-to-end way, they are known to be prone to picking up biases present in the training data. The dataset presented in this work contains a large number of patients compared to other datasets, which is expected to mitigate possible biases in the data. Still, all samples were obtained from a single center, so that sample processing and staining was done in the same way for all input data. Additionally, the dataset used in this work was collected using the same scanning hardware and data formats.

While this setup ensures that the dataset produced in this work is internally consistent,

enlargement of the data basis by including images obtained using different staining, illumination and scanning equipment would likely increase the generalisability of the networks' predictions. Likewise, inclusion of blood smears from other centers or re-training of the network with local data might help the model to generalise better or be adapted to the diagnostic environment of a different center. Compiling such a multi-sourced dataset and correcting for possible batch effects remains a challenge with a significant material and time cost.

Ultimately, a method of computer-aided diagnostics has to be tested in a routine application environment, in order to prospectively evaluate its performance in a prospective way. Scientific, regulatory and legal standards of the application of artificial intelligence models in a medical context are still rapidly evolving at the time of writing. Presently, a practicable application scenario seems to be using a network as a pre-sorting system that suggests classifications of leukocytes which then have to be validated by a human observer. Ultimately, networks could then be re-trained and learn from corrections made by the supervising human, although this would include the possibility to "learn mistakes", and is therefore not usually used in the medical context to guarantee model stability.

4.3 Perspectives

The present work led to two direct follow-up questions which have been worked on by the author, but are not included in this thesis. Firstly, the observation that the number of images present in the dataset is highly correlated with the success of network training leads to the question if there is an ideal number of training images to enable training of a successful single-cell classifier. At the time of writing of this thesis, no general answer to that question is known. However, dependence of classifier accuracy on sample number could be studied by systematically varying the size of the training set.

Secondly, it seems desirable to extend the approach taken in the present work to the morphology of bone marrow cells, whose examination typically follows upon appearance of a suspicious finding in the peripheral blood and represents the gold standard of diagnosis for many hematological disorders. However, as was the case so far for peripheral blood, publicly accessible databases for images of these cells are lacking at the time of writing.

In the present work, a dataset of over 18,000 single-cell leukocyte images was compiled and annotated, which formed the basis for developing neural networks that are able to classify leukocytes with a significantly better performance than previous approaches. The strategy used in this project was similar to the procedure often taken in natural image classification, where the software is trained to recover a label assigned by a human annotator. In the context of the present project, it was necessary to follow that strategy in particular due to the fact that the well-known morphological classes conventionally used in cytomorphologic reports are the product of long-running human classification efforts that do not perfectly

correlate with other diagnostic modalities. It is a drawback of this strategy that ultimately, there is no independent gold standard that could be used to decide the correctness of a cell classification independently of a human label. Therefore, the perfect algorithm is the one that reproduces the labels of the best available examiner, whose classification cannot be exceeded by definition.

This limitation could be overcome by compiling an imaging dataset that combines light-microscopic data as used in the context of the present work with labels obtained from additional, independent data, e.g. immunophenotypic or genetic information on a single-cell level. While building up such a database comes at a significantly higher cost, it would enable to make predictions from the microscopic data and go beyond the knowledge of a gold-standard annotator, potentially pointing towards correlations between the different diagnostic modalities that are hitherto unknown. For example, a network trained on such a dataset might be able to predict genetic properties of a cell from the microscopic image, and hence suggest associations between its genotype and phenotype, of which only relatively few are known today. Investigating the way in which networks make these predictions could point to subtle, so far unknown properties of light microscopic images. Neural networks as they are used today cannot themselves provide causal explanations of such phenomena, which remains the domain of human investigation. When trained on high-quality data, they can however be trained to become powerful aides, both in accurately emulating human tasks and pointing towards properties of the data that are difficult for humans to capture. The hope is that these properties will lead to a better understanding of human disease and a more reliable and accurate diagnostic procedures, which both will eventually work to the benefit of patients.

Appendix A

Annotation software

For the different stages of annotation in the present work, two independent annotation tools were developed, one for the gold-standard annotation of the whole AOI scan, and another one for re-annotation of single cell patches.

The first tool was developed based on JavaScript and jQuery, based on the open-source tool OpenSeadragon [120], a web-based viewer for display of high-resolution, zoomable images. It was designed to allow annotation of single leukocytes in an ergonomic and fast way, emulating the workflow of microscopic cell differentiation as closely as possible while building on experience with other commonly used applications, e.g. for the display of maps. For display with OpenSeadragon, scanned AOIs were first converted into the `.dzi` format. The tool shows the freely zoomable scanned AOI to the annotator, as well as the number of cells annotated so far on the right of the viewer region. Upon right mouse click into the display area, a dropdown menu offers the annotator all possible cell categories for annotation (cf. Fig. A.1). All annotations can be deleted and changed. When annotation is complete, all marked mouse positions are saved to an XML file, together with their pixel coordinates relative to the original image. These can then be used to extract the single-cell patches from the scanned AOI.

The second annotation tool was developed for efficient and ergonomic re-annotation of extracted single-cell images. Images are loaded from the re-annotation set in random order. The re-annotator can see the full-resolution single-cell image, and choose the morphological class of the image from a set of buttons displayed to the right of the image (cf. Fig. A.2). After completion of annotation, a full list of filename and re-annotation is output.

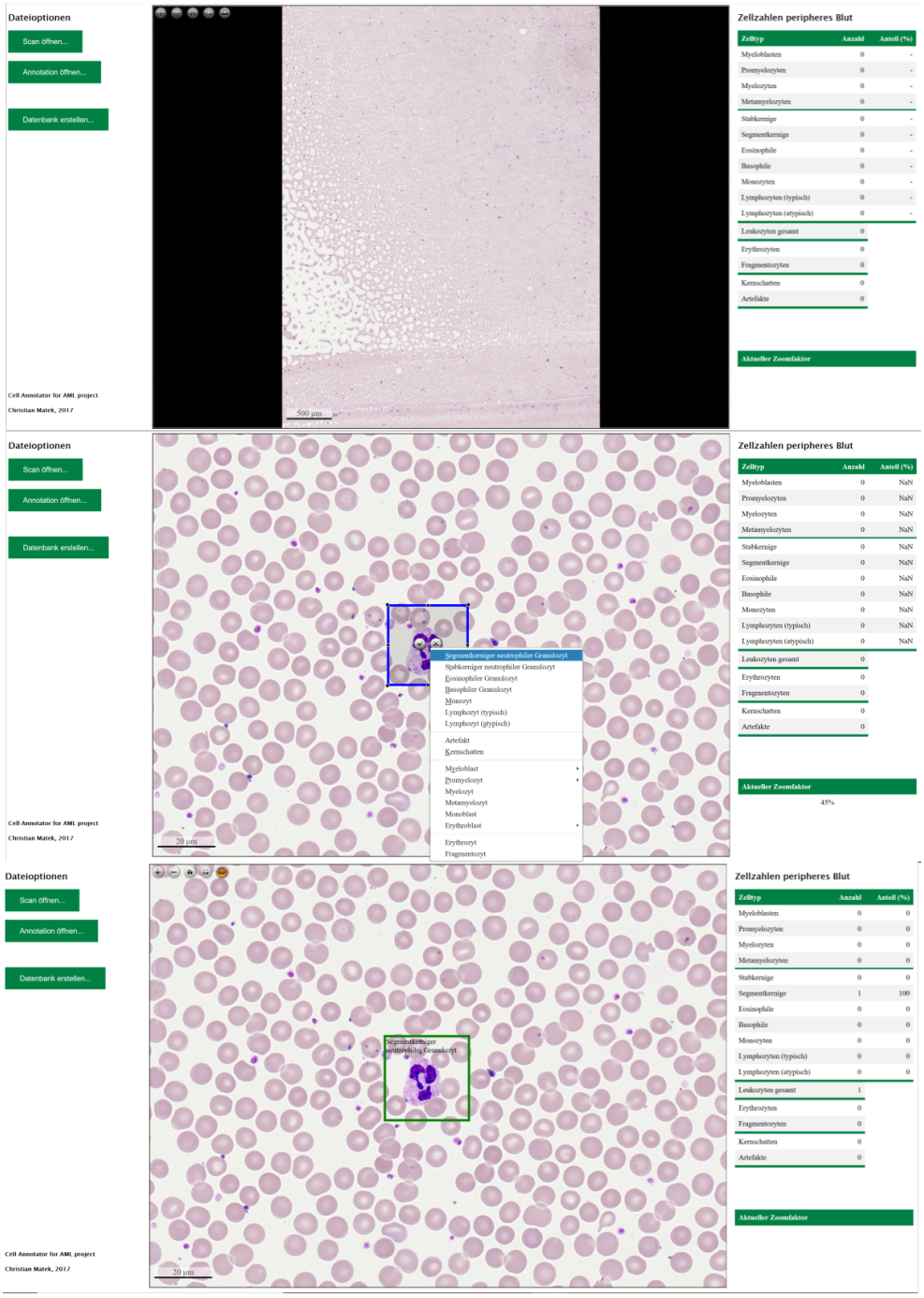


Figure A.1: Annotation of the scanned AOI using a deep-zoomable scan image.
Upper panel: Overview of the scanned AOI.
Middle panel: Selected cell area (blue bounding box) with dropdown selection of possible morphological classes.
Lower panel: Annotated single cell (green bounding box with annotated class visible).

Figure A.2: Software tool for re-annotations.
A single-cell image is displayed in full resolution, allowing the re-annotator to choose the morphological class from a set of buttons to the right.

Appendix B

Result for individual training folds

The results reported in Secs. 3.3 and 3.4 were obtained using 5-fold cross-validation. Hence, the corresponding confusion matrices were calculated by averaging across the individual confusion matrices of the 5 folds for which the ResNeXt and sequential networks were trained and tested. As detailed in Sec. 2.4.3, 80% of the overall single-cell image dataset was used for training, and the remaining 20% for testing the network trained for the respective fold. Test sets for the different folds are constructed to be mutually disjoint, so that each single-cell image is used for testing in only one of the 5 folds.

Confusion matrices of the individual folds are depicted in Fig. B.1 for the ResNeXt model, and in Fig. B.2 for the sequential model described in Sec. 2.4.2. Given within the individual entries of the confusion matrices is the number of images with the indicated classification behaviour.

For classes containing a sufficient number of single-cell images, sample noise is small and there is only minor variation between folds. Overall, the deviation of network prediction from the ground truth is similar to the deviation of a human examiner from the ground truth, as described in Sec. 3.2. As may be expected, inter-fold variability is higher for classes with small population due to sample noise.

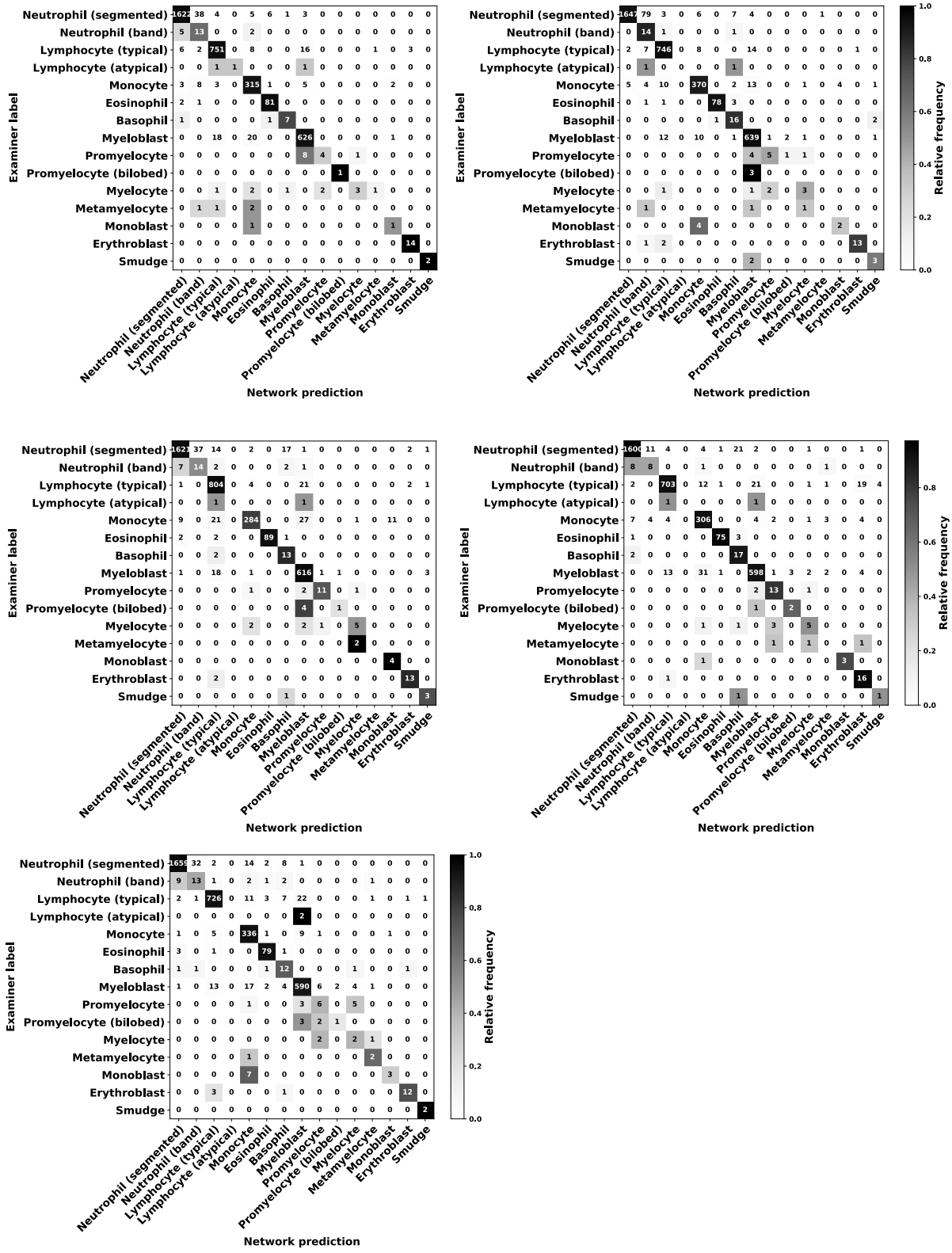


Figure B.1: Confusion matrices of individual folds used for 5-fold cross-validation of the ResNeXt network. Note that the prediction quality is good in all folds individually. Differences arise mostly for classes populated by few single-cell images through small size of test samples, corresponding to increased noise due to random sampling.

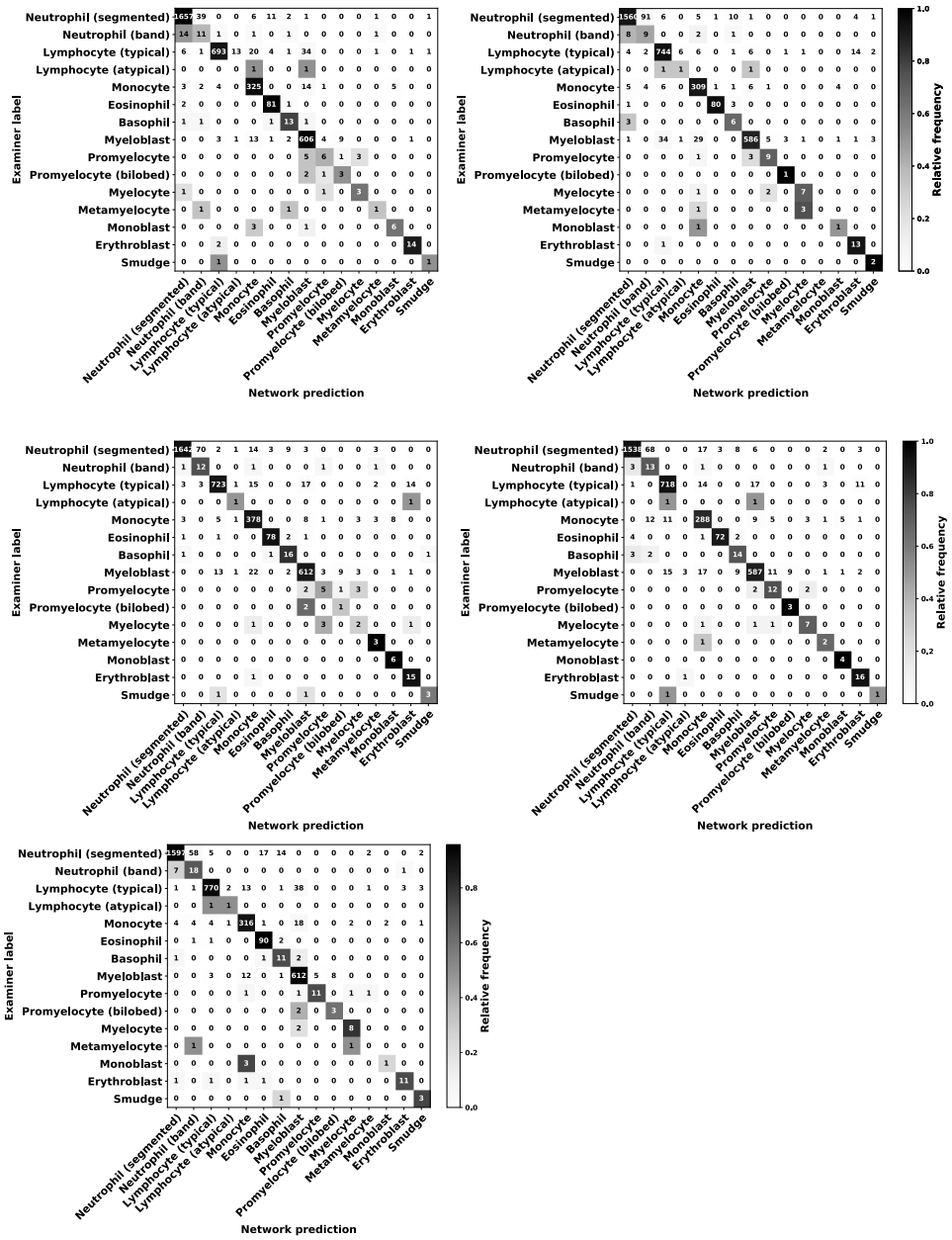


Figure B.2: Confusion matrices of individual folds used for 5-fold cross-validation of the sequential network. Also this network structure shows good prediction quality in all folds individually.

Appendix C

Structure of image dataset and code repository

The image dataset compiled in this work has been made publicly available through the TCIA database under `https://doi.org/10.7937/tcia.2019.36f5o9ld` [107].
For publication, images were sorted in a folder structure representing the ground-truth annotation provided by the first examiner. Within the morphological classes, the cell images were randomised, so that it cannot be inferred from file names which smear the cell images were taken from. Second and third annotations were additionally deposited in the file `annotations.dat`. A list of abbreviations used in the database is given in Tab. C.1.

abbreviation	description
BAS	Basophil
EBO	Erythroblast
EOS	Eosinophil
KSC	Smudge cell
LYA	Lymphocyte (atypical)
LYT	Lymphocyte (typical)
MMZ	Metamyelocyte
MOB	Monoblast
MON	Monocyte
MYB	Myelocyte
MYO	Myeloblast
NGB	Neutrophil (band)
NGS	Neutrophil (segmented)
PMB	Promyelocyte (bilobled)
PMO	Promyelocyte
UNC	Image that could not be assigned a class during re-annotation
nan	no re-annotation

Table C.1: Abbreviations used in TCIA deposition.

Bibliography

[1] J.H Bennett. Case of hypertrophy of the spleen and liver in which death took place from suppuration of the blood. *Edinburgh Medical and Surgical Journal*, 64:413 – 423, 1845.

[2] R. Virchow. Weisses Blut. *Froriep's Notizen*, 36:151 – 156, 1845.

[3] W. Ebstein. Über die acute Leukämie und Pseudoleukämie. *Deutsch Arch Klin Med.*, 44:343, 1889.

[4] O. Naegeli. Ueber rothes Knochenmark und Myeloblasten. *Deutsche Medizinische Wochenschrift.*, 26(18):287 – 290, 1900.

[5] E. M. Keohane, L. Smith, and J. M. Walenga. *Rodak's Hematology - Clinical Principles and Applications. Fifth edition.* Elsevier, 2016.

[6] A. Rad, M. Häggström, et al. Human hematopoiesis. *Wikipedia*, 2020.

[7] D. A. Arber, A. Orazi, R. Hasserjian, J. Thiele, M. J. Borowitz, M. M. Le Beau, C. D. Bloomfield, M. Cazzola, and J. W. Vardiman. The 2016 revision to the world health organization classification of myeloid neoplasms and acute leukemia. *Blood*, 127:2391–2405, 2016.

[8] I. De Kouchkovsky and M. Abdul-Hay. 'acute myeloid leukemia: a comprehensive review and 2016 update'. *Blood cancer journal*, 6:e441, 2016.

[9] T. Haferlach, U. Bacher, H. Theml, and H. Diem. *Taschenatlas Hämatologie, 6. Auflage.* Thieme, 2012.

[10] M. Wakui, K. Kuriyama, Y. Miyazaki, T. Hata, M. Taniwaki, S. Ohtake, H. Sakamaki, S. Miyawaki, T. Naoe, R. Ohno, and M. Tomonaga. Diagnosis of acute myeloid leukemia according to the WHO classification in the Japan Adult Leukemia Study Group AML-97 protocol. *International journal of hematology*, 87:144–151, 2008.

[11] E. F. Goljan. *Rapid review Pathology, fourth edition.* Elsevier, 2014.

[12] J. M. Bennett, D. Catovsky, M. T. Daniel, G. Flandrin, D. A. Galton, H. R. Gralnick, and C. Sultan. Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group. *British journal of haematology*, 33:451–458, 1976.

[13] J. M. Bennett, D. Catovsky, M. T. Daniel, G. Flandrin, D. A. Galton, H. R. Gralnick, and C. Sultan. Proposed revised criteria for the classification of Acute Myeloid Leukemia. A report of the French-American-British Cooperative Group. *Annals of internal medicine*, 103:620–625, 1985.

[14] C. Crump, J. Sundquist, W. Sieh, M. A. Winkleby, and K. Sundquist. Perinatal risk factors for acute myeloid leukemia. *European journal of epidemiology*, 30:1277–1285, 2015.

[15] G. Nagel, D. Weber, E. Fromm, S. Erhardt, M. Lübbert, W. Fiedler, T. Kindler, J. Krauter, P. Brossart, A. Kündgen, H. R. Salih, J. Westermann, G. Wulf, B. Hertenstein, M. Wattad, K. Götze, D. Kraemer, T. Heinicke, M. Girschikofsky, H. G. Derigs, H. A. Horst, C. Rudolph, M. Heuser, G. Göhring, V. Teleanu, L. Bullinger, F. Thol, V. I. Gaidzik, P. Paschka, K. Döhner, A. Ganser, H. Döhner, R. F. Schlenk, and German-Austrian AML Study Group (AMLSG). Epidemiological, genetic, and clinical characterization by age of newly diagnosed acute myeloid leukemia based on an academic population-based registry study (AMLSG BiO). *Annals of hematology*, 96:1993–2003, 2017.

[16] E. Hulegårdh, C. Nilsson, V. Lazarevic, H. Garelius, P. Antunovic, Å. Rangert Derolf, L. Möllgård, B. Uggla, L. Wennsträm, A. Wahlin, M. Höglund, G. Juliusson, D. Stockelberg, and S. Lehmann. Characterization and prognostic features of secondary acute myeloid leukemia in a population-based setting: a report from the Swedish Acute Leukemia Registry. *American journal of hematology*, 90:208–214, 2015.

[17] G. Juliusson, P. Antunovic, A. Derolf, S. Lehmann, L. Möllgård, D. Stockelberg, U. Tidefelt, A. Wahlin, and M. Höglund. Age and acute myeloid leukemia: real world data on decision to treat and outcomes from the Swedish Acute Leukemia Registry. *Blood*, 113:4179–4187, 2009.

[18] G. Piller. Leukaemia - a brief historical review from ancient times to 1950. *British journal of haematology*, 112:282–292, 2001.

[19] H. Dombret and C. Gardin. An update of current treatments for adult acute myeloid leukemia. *Blood*, 127:53–61, 2016.

[20] H. Döhner, E. Estey, D. Grimwade, S. Amadori, F. R. Appelbaum, T. Büchner, H. Dombret, B. L. Ebert, P. Fenaux, R. A. Larson, R. L. Levine, F. Lo-Coco, T. Naoe, D. Niederwieser, G. J. Ossenkoppele, M. Sanz, J. Sierra, M. S. Tallman, H. Tien,

A. H. Wei, B. Löwenberg, and C. D. Bloomfield. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood*, 129:424–447, 2017.

[21] A. Burnett, M. Wetzler, and B. Löwenberg. Therapeutic advances in acute myeloid leukemia. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 29:487–494, 2011.

[22] F. Lo-Coco, G. Avvisati, M. Vignetti, C. Thiede, S. M. Orlando, S. Iacobelli, F. Ferrara, P. Fazi, L. Cicconi, E. Di Bona, G. Specchia, S. Sica, M. Divona, A. Levis, W. Fiedler, E. Cerqui, M. Breccia, G. Fioritoni, H. R. Salih, M. Cazzola, L. Melillo, A. M. Carella, C. H. Brandts, E. Morra, M. von Lilienfeld-Toal, B. Hertenstein, M. Wattad, M. Lübbert, M. Hänel, N. Schmitz, H. Link, M. G. Kropp, A. Rambaldi, G. La Nasa, M. Luppi, F. Ciceri, O. Finizio, A. Venditti, F. Fabbiano, K. Döhner, M. Sauer, A. Ganser, S. Amadori, F. Mandelli, H. Döhner, G. Ehninger, R. F. Schlenk, U. Platzbecker, Gruppo Italiano Malattie Ematologiche dell'Adulto, German-Austrian Acute Myeloid Leukemia Study Group, and Study Alliance Leukemia. Retinoic acid and arsenic trioxide for acute promyelocytic leukemia. *The New England journal of medicine*, 369:111–121, 2013.

[23] N. J. Short, M. Konopleva, T. M. Kadia, G. Borthakur, F. Ravandi, C. D. DiNardo, and N. Daver. Advances in the Treatment of Acute Myeloid Leukemia: New Drugs and New Challenges. *Cancer discovery*, 2020.

[24] I. Lohse, K. Statz-Geary, S. P. Brothers, and C. Wahlestedt. Precision medicine in the treatment stratification of AML patients: challenges and progress. *Oncotarget*, 9:37790–37797, 2018.

[25] S. A. Patel and J. M. Gerber. A User's Guide to Novel Therapies for Acute Myeloid Leukemia. *Clinical lymphoma, myeloma & leukemia*, 2020.

[26] M. Heuser, A. Mina, E. M. Stein, and J. K. Altman. How Precision Medicine Is Changing Acute Myeloid Leukemia Therapy. *American Society of Clinical Oncology educational book. American Society of Clinical Oncology. Annual Meeting*, 39:411–420, 2019.

[27] J. H. Bennett. On Leucocythemia, or Blood Containing an Unusual Number of Colourless Corpuscles. *Mon. J. Med. Sci.*, 3:17–38, 1851.

[28] B. J. Bain. Diagnosis from the blood smear. *The New England journal of medicine*, 353:498–507, 2005.

[29] R. Hooke. *Lectures and Collections: Microscopium.* J. Martyn, London, 1678.

[30] A. van Leeuwenhoek. *Arcana Naturae Detecta.* H. van Kroonevelt, Delft, 1695.

[31] K.-A. Kreuzer, P. Bettelheim, T. Haferlach, and A. Rosenwald. Leitlinie Hämatologische Diagnostik. *Leitlinien der Deutschen Gesellschaft für Hämatologie und Onkologie*, 2018.

[32] S. Inoué and R. Oldenbourg. *Handbook of Optics, Volume II. Second Edition.*, chapter Microscopes, pages 17.1–17.52. McGraw - Hill, 1995.

[33] M. C. Montalto, R. R. McKay, and R. J. Filkins. Autofocus methods of whole slide imaging systems and the introduction of a second-generation independent dual sensor scanning method. *Journal of pathology informatics*, 2:44, 2011.

[34] Wikimedia, author ja:user:GcG. Oil immersion microscopy. Wikimedia commons, https://commons.wikimedia.org/wiki/File:Immersion_microscopy.jpg, 2020.

[35] Wikimedia, author Rollroboter. Herstellung eines blutausstriches. Wikimedia commons, https://commons.wikimedia.org/wiki/File:Blut_(1).jpg, 2020.

[36] M. Mulisch and U. Welsch, editors. *Romeis Mikroskopische Technik.* Springer Spektrum, Berlin, Heidelberg, 2015.

[37] A. Pappenheim. Zur Blutzellfärbung im klinischen Bluttrockenpräparat und zur histologischen Schnittpräparatfärbung der hämatopoetischen Gewebe nach meiner Methode. *Folia Haematol.*, 13:337–344, 1912.

[38] T. Binder, H. Diem, R. Fuchs, K. Gutensohn, and T. Nebe. Pappenheim Stain: Description of a hematological standard stain – history, chemistry, procedure, artifacts and problem solutions. *J. Lab. Med.*, 36:293–309, 2012.

[39] H. Löffler, J. Rastetter, and T. Haferlach. *Atlas of Clinical Hematology, 5th edition.* Springer-Verlag, Berlin Heidelberg, 2005.

[40] T. Haferlach and I. Schmidts. The power and potential of integrated diagnostics in acute myeloid leukaemia. *British journal of haematology*, 188:36–48, 2020.

[41] S. Kayser and M. J. Levis. Clinical implications of molecular markers in acute myeloid leukemia. *European journal of haematology*, 102:20–35, 2019.

[42] M. C. Béné, T. Nebe, P. Bettelheim, B. Buldini, H. Bumbea, W. Kern, F. Lacombe, P. Lemez, I. Marinov, E. Matutes, M. Maynadié, U. Oelschlagel, A. Orfao, R. Schabath, M. Solenthaler, G. Tschurtschenthaler, A. M. Vladareanu, G. Zini, G. C. Faure, and A. Porwit. Immunophenotyping of acute leukemia and lymphoproliferative disorders: a consensus proposal of the European LeukemiaNet Work Package 10. *Leukemia*, 25:567–574, 2011.

[43] F. Grignani, P. F. Ferrucci, U. Testa, G. Talamo, M. Fagioli, M. Alcalay, A. Mencarelli, F. Grignani, C. Peschle, and I. Nicoletti. The acute promyelocytic leukemia-specific PML-RAR alpha fusion protein inhibits differentiation and promotes survival of myeloid precursor cells. *Cell*, 74:423–431, 1993.

[44] G. J. Schuurhuis, M. Heuser, S. Freeman, M.-C. Béné, F. Buccisano, J. Cloos, D. Grimwade, T. Haferlach, R. K. Hills, C. S. Hourigan, J. L. Jorgensen, W. Kern, F. Lacombe, L. Maurillo, C. Preudhomme, B. A. van der Reijden, C. Thiede, A. Venditti, P. Vyas, B. L. Wood, R. B. Walter, K. Döhner, G. J. Roboz, and G. J. Ossenkoppele. Minimal/measurable residual disease in AML: a consensus document from the European LeukemiaNet MRD Working Party. *Blood*, 131:1275–1291, 2018.

[45] F. E. Craig and K. A. Foon. Flow cytometric immunophenotyping for hematologic neoplasms. *Blood*, 111:3941–3967, 2008.

[46] A. Lovelace. Notes upon L. F. Menabrea's "Sketch of the Analytical Engine invented by Charles Babbage". *Scientific Memoirs*, 3:691–731, 1842.

[47] C. Babbage. *Passages from the Life of a Philosopher*, chapter VIII - Of the Analytical Engine., pages 112 —- 141. Longman, Green, Longman, Roberts, & Green, 1864.

[48] A. Turing. Computing machinery and intelligence. *Mind*, LIX:433–460, 1950.

[49] D. O. Hebb. *The organization of behavior.* Wiley, New York, 1949.

[50] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65:386–408, 1958.

[51] D. H. Hubel. Evolution of ideas on the primary visual cortex, 1955-1978: a biased historical account (Nobel lecture). *Bioscience reports*, 2:435–469, 1982.

[52] M. Minsky and S. Papert. *Perceptrons: An Introduction to Computational Geometry. Second edition.* The MIT Press, Cambridge, MA, 1972.

[53] D. Crevier. *Ai: The Tumultuous History Of The Search For Artificial Intelligence.* Basic Books, 1993.

[54] P. J. Werbos. *The Roots of Backpropagation : From Ordered Derivatives to Neural Networks and Political Forecasting.* John Wiley & Sons, New York, 1994.

[55] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning.* MIT Press, 2016.

[56] J. Schmidhuber. Deep learning in neural networks: an overview. *Neural networks : the official journal of the International Neural Network Society*, 61:85–117, 2015.

[57] Wikimedia, author User:Dhp1080. Neuron description. Wikimedia commons, https://commons.wikimedia.org/wiki/Neuron/media/File:Neuron.svg, 2020.

[58] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18:1527–1554, 2006.

[59] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. nips 19 (pp. 153–160), 2007.

[60] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.*, 2012.

[61] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 16:411–418, 2013.

[62] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017.

[63] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, 316:2402–2410, 2016.

[64] D. Milea, R. P. Najjar, J. Zhubo, D. Ting, C. Vasseneix, X. Xu, M. Aghsaei Fard, P. Fonseca, K. Vanikieti, W. A. Lagrèze, C. La Morgia, C. Y. Cheung, S. Hamann, C. Chiquet, N. Sanda, H. Yang, L. J. Mejico, M.-B. Rougier, R. Kho, T. Thi Ha Chau, S. Singhal, P. Gohier, C. Clermont-Vignal, C.-Y. Cheng, J. B. Jonas, P. Yu-Wai-Man, C. L. Fraser, J. J. Chen, S. Ambika, N. R. Miller, Y. Liu, N. J. Newman, T. Y. Wong, V. Biousse, and BONSAI Group. Artificial Intelligence to Detect Papilledema from Ocular Fundus Photographs. *The New England Journal of Medicine*, 2020.

[65] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. C. Corrado, A. Darzi, M. Etemadi, F. Garcia-Vicente, F. J. Gilbert, M. Halling-Brown, D. Hassabis, S. Jansen, A. Karthikesalingam, C. J. Kelly, D. King, J. R. Ledsam, D. Melnick, H. Mostofi, L. Peng, J. J. Reicher, B. Romera-Paredes, R. Sidebottom, M. Suleyman, D. Tse, K. C. Young, J. De Fauw, and S. Shetty. International evaluation of an AI system for breast cancer screening. *Nature*, 577:89–94, 2020.

[66] M. Waldrop. News feature: What are the limits of deep learning? *Proc. Natl. Acad. Sci. U.S.A.*, 116:1074–1077, 2019.

[67] "MLAtlas". Imagenet performance history. https://paperswithcode.com/sota/image-classification-on-imagenet, 2020.

[68] J. Cho, K. Lee, E. Shin, G. Choy, and S. Do. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? *arXiv preprint arXiv:1511.06348*, 2015.

[69] Y. Gal, R. Islam, and Z. Ghahramani. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1183–1192. JMLR. org, 2017.

[70] C. Matek, S. Schwarz, K. Spiekermann, and C. Marr. A single-cell image database of leukocyte morphologies relevant in acute myeloid leukemia. *under review*, 2020.

[71] C. Matek, S. Schwarz, K. Spiekermann, and C. Marr. Human-level recognition of blast cells in acute myeloid leukemia with convolutional neural networks. *Nature Machine Intelligence*, (1):538–544, 2019.

[72] M. D. Zarella, D. Bowman, F. Aeffner, N. Farahani, A. Xthona, S. F. Absar, A. Parwani, M. Bui, and D. J. Hartman. A practical guide to whole slide imaging: A white paper from the digital pathology association. *Archives of Pathology & Laboratory Medicine.*, 143(2):222–234, 2019.

[73] F. Aeffner, H. A. Adissu, M. C. Boyle, R. D. Cardiff, E. Hagendorn, M. J. Hoenerhoff, R. Klopfleisch, S. Newbigging, D. Schaudien, O. Turner, et al. Digital microscopy, image analysis, and virtual slide repository. *ILAR journal*, 59(1):66–79, 2018.

[74] C. Shorten and T.M. Khoshgoftaar. A survey on image data augmentation for deep learning. *J Big Data*, 6(60), 2019.

[75] G. Al-Bdour, R. Al-Qurran, M. Al-Ayyoub, and A. Shatnawi. A detailed comparative study of open source deep learning frameworks. *arXiv preprint arXiv:1903.00102*, 2019.

[76] F. Seide and A. Agarwal. CNTK: Microsoft's Open-Source Deep-Learning Toolkit. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 2135, New York, NY, USA, 2016. Association for Computing Machinery.

[77] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, 2016.

[78] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, Vi. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015. Software available from tensorflow.org.

[79] F. Chollet et al. Keras 2.0. `https://keras.io`, 2017.

[80] W. Rawat and Z. Wang. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural computation*, 29:2352–2449, 2017.

[81] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for hyper-parameter optimization. In *Advances in neural information processing systems*, pages 2546–2554, 2011.

[82] E. Bochinski, T. Senst, and T. Sikora. Hyper-Parameter Optimization for Convolutional Neural Network Committees Based on Evolutionary Algorithms. 2017.

[83] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[84] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *arXiv:1611.05431 and CVPR2017*, 2016.

[85] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[86] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 448–456. JMLR.org, 2015.

[87] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi. A survey of the recent architectures of deep convolutional neural networks. *arXiv preprint arXiv:1901.06032*, 2019.

[88] R. K. Srivastava, K. Greff, and J. Schmidhuber. Training Very Deep Networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2377–2385. Curran Associates, Inc., 2015.

[89] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[90] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[91] S. Zagoruyko and N. Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, 2016.

[92] M. Dietz. ResNeXt implementation for Keras. https://gist.githubusercontent.com/mjdietzx/, 2017.

[93] G. Eraslan, Ž. Avsec, J. Gagneur, and F. J. Theis. Deep learning: new computational modelling techniques for genomics. *Nature reviews. Genetics*, 20:389–403, 2019.

[94] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv preprintarXiv:1412.6980*, 2014.

[95] D. Bychkov, N. Linder, R. Turkki, S. Nordling, P. E. Kovanen, C. Verrill, M. Walliander, M. Lundin, C. Haglund, and J. Lundin. Deep learning based tissue analysis predicts outcome in colorectal cancer. *Scientific reports*, 8:3395, 2018.

[96] R. Geirhos, D. H. J. Janssen, H. H. Schütt, J. Rauber, M. Bethge, and F. A. Wichmann. Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969v2*, 2017.

[97] R. Geirhos, C. R. Medina Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann. Generalisation in humans and deep neural networks. *arXiv preprint arXiv:1808.08750v2*, 2018.

[98] Z. Zhang, M. W. Beck, D. A. Winkler, B. Huang, W. Sibanda, H. Goyal, and written on behalf of AME Big-Data Clinical Trial Collaborative Group. Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Annals of translational medicine*, 6:216, 2018.

[99] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller. Causability and explainability of artificial intelligence in medicine. *Wiley interdisciplinary reviews. Data mining and knowledge discovery*, 9:e1312, 2019.

[100] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[101] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[102] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.

[103] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[104] A. Shrikumar, P. Greenside, and A. Kundaje. Learning Important Features through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3145–3153. JMLR.org, 2017.

[105] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PloS one*, 10:e0130140, 2015.

[106] D. S. Char, N. H. Shah, and D. Magnus. Implementing Machine Learning in Health Care - Addressing Ethical Challenges. *The New England journal of medicine*, 378:981–983, March 2018.

[107] C. Matek, S. Schwarz, C. Marr, and K. Spiekermann. A Single-cell Morphological Dataset of Leukocytes from AML Patients and Non-malignant Controls. *The Cancer Imaging Archive https://doi.org/10.7937/tcia.2019.36f5o9ld*, 2019.

[108] S. Krappe, M. Benz, T. Wittenberg, T. Haferlach, and C. Münzenmayer. Automated classification of bone marrow cells in microscopic images for diagnosis of leukemia: a comparison of two classification schemes with respect to the segmentation quality. In Lubomir M. Hadjiiski and Georgia D. Tourassi, editors, *Medical Imaging 2015: Computer-Aided Diagnosis*, volume 9414, pages 858 – 863. International Society for Optics and Photonics, SPIE, 2015.

[109] J. L. Fleiss, B. Levin, and M. Cho Paik. *Statistical Methods for Rates and Proportions*. Wiley, 2003.

[110] X. Fuentes-Arderiu and D. Dot-Bach. Measurement uncertainty in manual differential leukocyte counting. *Clinical chemistry and laboratory medicine*, 47:112–115, 2009.

[111] S. Krappe, T. Wittenberg, T. Haferlach, and C. Münzenmayer. Automated morphological analysis of bone marrow cells in microscopic images for diagnosis of leukemia: nucleus-plasma separation and cell classification using a hierarchical tree model of hematopoesis. Bildverarbeitung für die Medizin 2016 : Algorithmen - Systeme - Anwendungen; Proceedings des Workshops vom 13. bis 15. März 2016 in Berlin, 2016.

[112] J. N. Mandrekar. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316, 2010.

[113] D.W. Hosmer and S. Lemeshow. *Applied Logistic Regression, Second Edition.* John Wiley & Sons, Inc., 2000.

[114] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K. Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 2017.

[115] J. W. Choi, Y. Ku, B. W. Yoo, J.-A. Kim, D. S. Lee, Y. J. Chai, H.-J. Kong, and H. C. Kim. White blood cell differential count of maturation stages in bone marrow smear using dual-stage convolutional neural networks. *PloS one*, 12:e0189259, 2017.

[116] F. Xing and L. Yang. Robust Nucleus/Cell Detection and Segmentation in Digital Pathology and Microscopy Images: A Comprehensive Review. *IEEE Rev Biomed Eng.*, 9(234-263), 2016.

[117] E. Cuevas, D. Oliva, M. Díaz, D. Zaldivar, M. Pérez-Cisneros, and G. Pajares. White Blood Cell Segmentation by Circle Detection Using Electromagnetism-Like Optimization. *Computational and Mathematical Methods in Medicine*, ID 395071, 2013.

[118] Y. M. Alomari, S. N. H. Sheikh Abdullah, R. Zaharatul Azma, and K. Omar. Automatic Detection and Quantification of WBCs and RBCs using Iterative Structured Circle Detection Algorithm. *Computational and Mathematical Methods in Medicine*, ID 979302, 2014.

[119] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017.

[120] "OpenSeadragon contributors". Openseadragon 2.2.1. `http://openseadragon.github.io`, 2016.

Acknowledgements

I would like to thank my co-supervisors Prof. Dr. Karsten Spiekermann and Dr. Carsten Marr for their constant support, interest and open-mindedness, allowing me to pursue a line of research bridging AI-based bioinformatics and clinical medicine.
Thanks are also due to the heads of the two academic institutions involved, Prof. Dr. Dr. Michael von Bergwelt and Prof. Dr. Dr. Fabian Theis for creating an environment open to interdisciplinary research.

At the Department of Medicine III in Großhadern, I am especially grateful to Simone Schwarz and Antje Holzäpfel for agreeing to take over the annotation task, without which the present project would not have been possible. Sebastian Tschuri provided help with access to samples and data structures. I also thank PD Dr. Klaus Metzeler for helpful discussions. To all other members of the Laboratory of Leukemia Diagnostics, I am grateful for welcoming me into their team.

At the Institute of Computational Biology, I am grateful to Prof. Dr. Dr. Fabian Theis for agreeing to be part of my Thesis Advisory Committee. I benefited from numerous discussions with the members of ICB, especially with Dr. Nikos Chlis and Dr. Tingying Peng on the subtleties of deep learning methods.

At Precipoint in Freising, I would like to thank all employees for their support during the digitization process of the blood smear samples, in particular Klaus Rattenhuber for technical guidance in using the microscopic equipment. I could discuss many aspects of digital pathology with Dominik Gerber, Helmut Krönauer, Friedrich Müller and Nikolas Weiß.

Being part of CRC 1243 during this project, I benefited from discussions with members of that group, which allowed for regular excursions into a wide range of research in oncology beyond my own project. The CRC organizers, especially Elisabeth Schröder-Reiter and Elke Hammerbacher, enabled this experience through their constant support.

It is a pleasure to thank Prof. Dr. Dr. Torsten Haferlach for discussions on many aspects of hematological diagnostics and its digitization.

This work was enabled by funding of Deutsche Forschungsgemeinschaft through CRC 1243,

and by a dissertation scholarship of Deutsche José Carreras Leukämie-Stiftung, for which I am very grateful.

For support during the work on this thesis and much more, I thank my family.