Semi-Supervised Training with Fuzzy Logic for Transformer-Based Semantic Segmentation

mgr inż. Michał Wieczorek

Doctoral dissertation supervised by
prof. dr hab. inż. Marcin Woźniak

Discipline: Information and Communication Technology

Department of Artificial Intelligence Modeling
Silesian University of Technology
Poland
2026

Abstract

This dissertation investigates how to design artificial-intelligence systems for medical and biologically relevant image analysis that remain accurate, data-efficient, computationally feasible, and reliable in the presence of annotation ambiguity. The work is structured as a cumulative research program that spans custom neural architectures for classification, semantic segmentation, synthetic data generation, and fuzzy uncertainty modeling. These studies culminate in a semi-supervised Transformer framework for hematological semantic segmentation that combines stain-aware preprocessing, pooled-memory global attention, spectral compression, masked supervision, and RGB robust fuzzy C-means regularization within a single training pipeline. Across the dissertation, the results show that domain-adapted architectures and principled data-efficiency strategies can outperform generic baselines in small-data scientific settings, while fuzzy modeling improves behavior on ambiguous borders and weakly annotated regions. The final hematological system achieves 99.73% validation accuracy, 99.73% micro-Dice, and 99.46% micro-IoU in a low-label regime, supporting the thesis that efficient long-range context modeling and uncertainty-aware semi-supervision can be integrated into a practical framework for blood-smear microscopy.

Keywords: medical image analysis, semantic segmentation, transformer, semi-supervised learning, fuzzy logic, hematological microscopy

Streszczenie

Niniejsza rozprawa doktorska bada metody projektowania systemów sztucznej inteligencji do analizy obrazów medycznych i biologicznych, które zachowują wysoką dokładność, efektywność danych, wykonalność obliczeniową oraz niezawodność w warunkach niejednoznaczności adnotacji. Praca została sformułowana jako kumulatywny program badawczy obejmujący dedykowane architektury neuronowe do klasyfikacji, segmentacji semantycznej, generowania danych syntetycznych oraz modelowania niepewności rozmytej. Badania te wieńczy opracowanie półnadzorowanego szkieletu opartego na Transformerach do segmentacji hematologicznej, który integruje w jednym procesie uczącym przetwarzanie wstępne uwzględniające specyfikę barwienia, globalną atencję z pulowaną pamięcią, kompresję spektralną, maskowany nadzór oraz regularyzację opartą na odpornych algorytmach rozmytych C-średnich w przestrzeni RGB. Wyniki przedstawione w rozprawie dowodzą, że architektury adaptowane domenowo oraz pryncypialne strategie efektywności danych przewyższają ogólne modele bazowe w naukowych zastosowaniach niskodanych, podczas gdy modelowanie rozmyte poprawia działanie systemu na niejednoznacznych krawędziach i obszarach słabo adnotowanych. Finalny system hematologiczny osiąga dokładność walidacyjną na poziomie 99,73%, wskaźnik micro-Dice równy 99,73% oraz micro-IoU wynoszący 99,46% w reżimie niskiego zapotrzebowania na etykiety, co potwierdza tezę, że efektywne modelowanie kontekstu dalekiego zasięgu oraz uwzględniający niepewność półnadzór mogą zostać zintegrowane w praktyczne narzędzie do mikroskopii wymazów krwi.

Slowa kluczowe: analiza obrazow medycznych, segmentacja semantyczna, transformatory, uczenie polnadzorowane, logika rozmyta, mikroskopia hematologiczna

Theses

In this dissertation, I developed a semi-supervised Transformer-based framework for hematological semantic segmentation of blood-smear microscopy images affected by limited annotation, staining variability, overlapping cells, and ambiguous object boundaries. In this work, the following theses have been established:

i. **Development of latent-compression attention mechanisms for image processing can improve the generalization capabilities of Transformer-based semantic segmentation for animal blood-sample analysis.**

ii. **A fuzzy-based semi-supervised training algorithm for medical image segmentation can improve pixel classification under ambiguous data conditions.**

iii. **Introducing fuzzy logic techniques into semantic-segmentation architectures can reduce pixel-level classification errors on ambiguous data and object borders.**

The framework combines quantile-based stain and background normalization with paired augmentation and cached tensor processing. Its SegFormer-inspired encoder-decoder is enhanced with HybridGlobalAttention, which uses pooled global memory, spectral compression, FFT-domain filtering, decimation, and a gated local branch to capture global smear context while preserving cellular details.

The training pipeline uses mixed labeled and unlabeled batches. Labeled samples are optimized with a masked supervised fuzzy loss, while all images contribute through an RGB RFCM+ unsupervised regularization loss that forms coherent fuzzy pixel prototypes and penalizes spatial inconsistency.

Overall, the model reduces dependence on dense manual annotation, improves efficient long-range context modeling, and handles uncertain blood-cell borders through fuzzy membership instead of premature hard labeling. Exact details om proposed model construction are presented in Section 9.

Contents

Abstract i

Streszczenie ii

Theses iii

1 Introduction . . . 1
- 1.1 AI in medical applications: a historical perspective . . . 1
- 1.2 Deep learning for tissue classification in microscopy . . . 1
- 1.3 Semantic segmentation in MRI and hematological imaging . . . 2
- 1.4 Fuzzy logic for handling ambiguous data . . . 2
- 1.5 Towards semi-supervised semantic segmentation with fuzzy logic in blood analysis . . . 3

2 Glossary . . . 4
- 2.1 Core fields, data, and research settings . . . 4
- 2.2 Neural networks and internal representations . . . 6
- 2.3 How neural networks learn . . . 7
- 2.4 Architectures for images and sequences . . . 8
- 2.5 Attention and transformer terminology . . . 10
- 2.6 Medical imaging and biomedical observation . . . 11
- 2.7 Fuzzy logic, uncertainty, and ambiguous boundaries . . . 12
- 2.8 Evaluation, reliability, and experimental methodology . . . 13
- 2.9 Signal processing and efficient computation . . . 14

3 Theoretical foundations of artificial intelligence . . . 16
- 3.1 Mathematical and computational preliminaries . . . 16
- 3.2 The artificial neuron and activation functions . . . 16
- 3.3 Artificial neural networks (ANNs) and deep learning . . . 16
- 3.4 The backpropagation algorithm . . . 17
- 3.5 Convolutional neural networks (CNNs) . . . 17
- 3.6 The attention mechanism . . . 17
- 3.7 The transformer architecture . . . 18
- 3.8 Vision transformers (ViT) . . . 18

3.9 Advanced optimization and regularization 18
4 Malaria detection using deep learning 19
4.1 Introduction and comprehensive background 19
4.2 Epidemiology, parasitology, and clinical significance 21
4.3 Foundations of deep learning and semantic segmentation . . . 23
4.4 Dataset curation and advanced preprocessing pipeline 23
4.5 Custom semantic segmentation neural network architecture . . 26
4.6 Experimental results and performance evaluation 28
4.7 Discussion, limitations, and future research trajectories 31
4.8 Conclusion 32
5 Transformer-based semantic segmentation for medical imaging 34
5.1 Introduction 34
5.2 Related works 35
5.3 Theoretical foundations: from self-attention to vision transformers 36
5.4 Dataset and preprocessing 38
5.5 Proposed architecture: the transformer-based semantic segmentation network (TSSN) 40
5.6 Training procedure 41
5.7 Experimental results 41
5.8 Discussion and limitations 45
5.9 Conclusion 45
6 Semantic segmentation for automatic weapon detection 46
6.1 Introduction and background 46
6.2 Related works in weapon detection and synthetic data 48
6.3 Synthetic dataset generation: a paradigm of control, scale, and fidelity 50
6.4 Neural network architecture: a deep dive into design and mathematics 54
6.5 Training methodology: optimization, loss, and regularization . 55
6.6 Experimental results and comprehensive analysis 56
6.7 Limitations, ethical considerations, and future work 60
6.8 Conclusion 61
7 Type-2 fuzzy client reputation for federated learning 63
7.1 Introduction and theoretical foundations 63
7.2 Background: federated learning architectures and security landscape 65
7.3 Mathematical foundations of type-2 fuzzy logic 67
7.4 Proposed client reputation mechanism: complete formalization 68

7.5 Experimental evaluation and results analysis 70
7.6 Conclusion 73
8 Custom deep learning with image augmentation for gastrointestinal tissue classification 75
8.1 Motivation, scope and scientific context 75
8.2 Dataset acquisition, curation and taxonomic organization . . . 76
8.3 Mathematical pre-processing and normalisation 78
8.4 Comprehensive data augmentation strategy 78
8.5 Network architecture: design rationale, detailed specification and theoretical foundations 80
8.6 Training procedure, optimisation strategies and hyper-parameter tuning 81
8.7 Evaluation protocols, metrics and empirical results 83
8.8 Discussion of methodological choices, limitations and potential biases 86
8.9 Concluding remarks 87
9 Semi-supervised transformer framework for hematological semantic segmentation 90
9.1 Introduction and motivation 90
9.2 Position of the final framework within the thesis 91
9.3 Clinical and technical need for a semi-supervised hematological system 91
9.4 Related works 92
9.5 Problem formulation and data regime 95
9.6 Methodology: design and implementation of the semi-supervised transformer framework 96
9.7 Evaluation protocol 106
9.8 Comparative evaluation strategy 108
9.9 Section summary 113
10 Integrated discussion 122
10.1 The self-attention bottleneck in high-resolution microscopy . . 122
10.2 Morphological variability and staining artifacts 122
10.3 Extending fuzzy logic to deep semantic maps 123
10.4 Architectural synthesis of the training workflow 123
10.5 Hardware considerations and inference optimization 123
10.6 Statistical results and ablation studies 124
11 Conclusion 125
11.1 Verification of the dissertation theses 125
11.2 Scientific contributions of the dissertation 127

11.3 Final assessment and future research 128

References **130**

1 Introduction

Artificial Intelligence (AI) has increasingly become a cornerstone of modern medical applications because it can enhance diagnostic accuracy, support treatment planning, and improve the consistency of quantitative image analysis [1], [2], [3]. With the integration of machine learning and deep learning algorithms, AI systems have demonstrated capabilities comparable to, and in some cases surpassing, those of human experts in tightly defined diagnostic tasks [2], [4]. In this thesis, I investigate how to design robust and data-efficient deep learning systems for medical image understanding, with particular emphasis on semantic segmentation in hematological microscopy.

I begin with foundational questions concerning architectural adaptation, feature learning, and optimization in medical images. I then examine dense prediction, synthetic data generation, and explicit uncertainty modeling with fuzzy logic. Each stage addresses one limitation that prevented the direct construction of the final system. In this way, the later semi-supervised Transformer for blood-smear segmentation emerges as the logical consequence of the earlier studies rather than as an isolated final chapter.

1.1 AI in medical applications: a historical perspective

The application of AI in medicine dates back to the 1970s with the advent of rule-based expert systems such as MYCIN [5], which was designed for infectious disease diagnosis. However, limitations in computational power and data availability curtailed the widespread adoption of early systems. The resurgence of AI in the 2010s, driven by the success of deep learning and the proliferation of large-scale annotated datasets, has transformed the field [1], [2]. Convolutional Neural Networks (CNNs), in particular, catalyzed breakthroughs in image recognition tasks, including radiography, histopathology, and endoscopy [2], [3], [6]. Today, AI assists clinicians in identifying tumors, segmenting organs, and predicting disease progression with unprecedented precision [2], [4].

1.2 Deep learning for tissue classification in microscopy

Human tissue classification from microscopy images represents one of the most challenging and essential tasks in medical diagnostics. Traditionally, CNNs have served as the backbone of such classification systems, extracting hierarchical spatial features through convolutional filters [1], [6]. While CNNs excel at capturing local texture and morphology, their inherent limitation lies in their restricted receptive fields and locality bias [1], [7].

The emergence of Transformer-based architectures, originally developed for natural language processing, has substantially influenced computer vision through models such as the Vision Transformer (ViT) [7], [8]. By employing self-attention mechanisms, transformers effectively capture long-range dependencies and global contextual information within tissue samples. Transformer-based models have therefore become an important design option for medical imaging tasks in which global context complements local texture cues [7], [9], [10].

1.3 Semantic segmentation in MRI and hematological imaging

Semantic segmentation, the process of assigning a class label to each pixel in an image, is critical in medical image analysis for tasks such as lesion delineation, organ boundary detection, and cell counting [2], [11], [12]. In Magnetic Resonance Imaging (MRI), segmentation facilitates quantitative analysis of brain structures, tumor volumes, and vascular abnormalities. Similarly, in hematological imaging, segmentation aids in isolating blood components, identifying abnormal cell morphologies, and tracking disease progression in disorders like leukemia [2], [3].

Modern segmentation models leverage encoder-decoder architectures such as U-Net and its variants, which combine high-resolution spatial information with deep semantic understanding [12], [13]. These architectures have been widely adopted due to their performance on limited datasets and their suitability for biomedical image dimensions. More recent innovations integrate attention mechanisms, multi-scale feature aggregation, and hybrid CNN-transformer layers to further enhance segmentation accuracy [9], [13], [14], [15], [16].

1.4 Fuzzy logic for handling ambiguous data

Medical data often contains ambiguity and imprecision due to factors such as overlapping cell structures, staining artifacts, and inter-observer variability. Fuzzy logic, with its ability to model uncertainty and partial truth, provides a valuable framework for dealing with such complexities [17], [18], [19]. Unlike crisp classification systems that assign a single label per input, fuzzy logic assigns degrees of membership to multiple classes, reflecting the inherent ambiguity of real-world medical data.

Incorporating fuzzy inference systems into AI models enables more nuanced decision-making and improves performance in borderline or noisy cases. Moreover, fuzzy clustering methods, such as Fuzzy C-Means, facilitate more flexible segmentation and grouping of image regions with gradual intensity transitions [19], [20].

1.5 Towards semi-supervised semantic segmentation with fuzzy logic in blood analysis

Despite significant advancements in supervised learning, the performance of segmentation models remains highly dependent on large volumes of annotated data. In the context of blood sample analysis, obtaining pixel-level annotations is time-consuming and requires domain expertise. Semi-supervised learning (SSL) offers a promising alternative by leveraging unlabeled data alongside limited labeled examples to enhance model generalization [21], [22].

A novel approach involves the integration of fuzzy logic into semi-supervised semantic segmentation frameworks. During training, fuzzy memberships can be used to refine pseudo-labels generated on unlabeled data, allowing the model to express uncertainty and avoid overconfident predictions on ambiguous regions. This fuzzy-SSL paradigm is particularly effective in blood smear analysis, where morphological overlap between different cell types introduces significant classification challenges [23], [24].

Preliminary experiments suggest that incorporating fuzzy membership degrees in the loss function and label refinement stages leads to improved segmentation performance and better model calibration. This hybrid approach opens new avenues for robust, data-efficient medical AI systems capable of handling real-world uncertainty in clinical practice.

2 Glossary

In this thesis, I use concepts from artificial intelligence, computer vision, signal processing, mathematics, and medical image analysis. Because many later chapters become highly technical, I try to explain here the most basic terms.

2.1 Core fields, data, and research settings

The terminology in this subsection follows standard modern treatments of deep learning, transfer learning, federated learning, and data-centric computer vision [1], [25], [26], [27], [28].

Artificial intelligence (AI) is the broad scientific and engineering field concerned with building systems that can perform tasks that usually require human intelligence, such as perception, reasoning, learning, and decision-making.

Computer vision is the branch of AI that teaches computers how to interpret images and videos. In this thesis, it is the main field through which I study medical and biological images.

Natural language processing (NLP) is the field of AI devoted to language, including text and speech. It is important here because the Transformer architecture was first developed for language tasks before being adapted to images.

Machine learning (ML) is a subfield of AI in which a system improves its performance by learning patterns from data instead of following only fixed hand-written rules.

Deep learning (DL) is a branch of machine learning based on neural networks with many layers. It is especially effective when the data are complex and high-dimensional, such as images, text, sound, or biological signals.

An algorithm is a precise step-by-step procedure for solving a problem. A learning algorithm tells a model how to improve itself from data.

A model is the mathematical system that produces predictions. In practical terms, it is the trainable function that receives an input and returns an output, such as a class label or a segmentation mask.

A dataset is a structured collection of examples used for analysis or training. In image analysis, a dataset usually contains images together with labels, annotations, or metadata.

A sample (or instance) is one individual element of a dataset. A sample may be a single image, one image patch, one patient record, or one sequence.

A feature is any measurable property of the data that may help the model solve the task. In classical machine learning, features are often designed by hand. In deep learning, the model usually learns useful features automatically.

A class is one of the categories that the model is expected to recognize, such as background, erythrocyte, leukocyte, or thrombocyte.

A label is the target answer assigned to a sample. For classification, it is usually one class per image. For segmentation, it may be one class per pixel.

Annotation is the process of assigning labels to data, usually by an expert or trained operator. In medical imaging, annotation can be slow and expensive because it often requires specialist knowledge.

Ground truth is the reference annotation treated as the best available correct answer. It is called "ground truth" even though, in many real medical settings, it still contains some uncertainty.

A training set is the portion of the data used to teach the model.

A validation set is a separate portion of the data used during development to compare design choices and tune settings without directly training on those examples.

A test set is the final held-out portion of the data used to estimate how well the finished model performs on previously unseen examples.

Supervised learning is the learning setting in which the model is trained using inputs together with correct target outputs.

Unsupervised learning is the learning setting in which the model works without explicit target labels and instead discovers structure, patterns, or groups directly from the data.

Semi-supervised learning (SSL) combines a small amount of labeled data with a larger amount of unlabeled data. This setting is especially important in medicine, where labels are valuable but difficult to obtain.

Transfer learning means starting from a model that has already learned from one task or dataset and adapting it to a new but related task. This is useful when the new dataset is too small to train a strong model from the beginning.

Fine-tuning is the practical process of continuing the training of a previously learned model on a new target dataset. It is one of the main ways transfer learning is implemented in practice.

Federated learning is a training framework in which multiple devices or institutions collaborate to improve a shared model without sending all raw data to one central location. This is valuable when privacy, regulation, or data ownership make direct data sharing difficult.

Synthetic data are artificial examples produced by a simulator, rendering pipeline, or generative process instead of being collected directly from the real world. Synthetic data are often used when real data are scarce, sensitive, or expensive.

Domain randomization is a strategy for creating synthetic data in which many aspects of the simulated scene are deliberately varied, such as lighting, color, texture, pose, or noise. The purpose is to make the model robust enough to work on real

data later.

Data augmentation means creating modified versions of existing training examples, for example by rotation, flipping, scaling, color changes, blur, or noise. Augmentation helps the model learn robustness without requiring entirely new data collection.

2.2 Neural networks and internal representations

The concepts in this subsection are grounded in the classical and modern neural network literature [1], [29], [30].

Artificial neural networks (ANNs) are computational models inspired loosely by biological nervous systems. They are built from many simple processing units connected into layers.

A neuron (or node) is a basic processing unit inside a neural network. It receives numerical inputs, combines them, and produces an output.

A layer is a group of neurons operating at the same stage of the model. Early layers usually capture simpler patterns, while deeper layers often represent more abstract information.

A weight is a trainable number that controls how strongly one signal influences another inside the network.

A bias is an additional trainable value added to a neuron before the activation function is applied. It gives the model extra flexibility.

An activation function is the mathematical rule that transforms the output of a neuron. Activation functions introduce non-linearity, which is essential if the network is to learn complex relationships instead of only simple straight-line ones.

A tensor is the standard way of storing numerical data in deep learning. A single number is a scalar, a one-dimensional list is a vector, a two-dimensional table is a matrix, and higher-dimensional arrays are called tensors.

A feature map is the structured output of a convolutional layer. It shows where a particular learned pattern appears in the image.

An embedding is a learned numerical representation of data in a form that is useful for a model. The purpose of an embedding is to convert raw inputs into a representation where important relationships become easier to learn.

A latent representation is an internal representation learned by the model that is usually not directly visible to the user. It is called "latent" because it captures hidden structure that explains the data.

A logit is the raw numerical score produced by a model before it is converted into a probability.

Softmax is the mathematical operation that converts a set of logits into a

probability distribution whose values sum to one. It is commonly used when the model must choose among multiple classes.

2.3 How neural networks learn

The optimization language used in this subsection follows standard backpropagation and adaptive-optimization references [31], [32], [33].

Training is the process through which a model adjusts its internal parameters so that its predictions become closer to the desired outputs.

Inference is the stage in which a trained model is used to make predictions on new data. Training is for learning; inference is for application.

A loss function is the numerical measure of how wrong the model currently is. Lower loss means that the prediction is closer to the desired answer.

Cross-entropy loss is a common loss function for classification. It penalizes the model when it assigns low probability to the correct class.

Dice loss is a loss function derived from the Dice overlap measure. It is commonly used in segmentation because it directly rewards overlap between the predicted region and the target region.

An optimizer is the rule used to update the model parameters during training. Popular optimizers include Stochastic Gradient Descent (SGD) and Adam.

Stochastic gradient descent (SGD) is a simple and widely used optimizer that updates parameters using gradient information from small batches of data.

Adam is an optimizer that adapts the update size for each parameter by using running estimates of past gradients. It often converges faster than basic SGD in practical deep learning tasks.

A gradient is information about how a small change in a parameter would change the loss. It tells the optimizer which direction is likely to improve the model.

Gradient descent is the basic idea of moving the model parameters in the direction that reduces the loss.

Backpropagation is the procedure used to compute gradients efficiently through all layers of a neural network. It works by propagating error information backward from the output to the earlier layers.

A batch is a small group of training examples processed together in one update step.

An epoch means one full pass through the entire training set.

An iteration is one parameter update step, usually based on one batch. Several iterations are needed to complete one epoch.

A hyperparameter is a setting chosen by the researcher before or during training, rather than learned automatically by the model. Examples include the

learning rate, batch size, and number of layers.

The learning rate controls how large each parameter update is during training. If it is too small, learning may be very slow. If it is too large, training may become unstable.

A learning-rate schedule is a rule for changing the learning rate over time. This is often done to make training stable at the beginning and precise near the end.

Warmup is a training strategy in which the learning rate starts small and gradually increases during the first part of training. This often helps very large models train more safely.

Regularization refers to techniques that reduce the risk of the model memorizing the training data too narrowly. The purpose is to improve performance on new, unseen examples.

Weight decay is a regularization method that discourages weights from becoming unnecessarily large. It often improves stability and generalization.

Overfitting occurs when the model learns the training data too specifically, including noise and accidental details, and therefore performs poorly on new data.

Underfitting occurs when the model is too simple or insufficiently trained to capture the main patterns even in the training data.

Generalization is the ability of a model to perform well on new data that were not used during training. In scientific and medical applications, good generalization is more important than merely fitting the training set.

Class imbalance means that some classes appear much more often than others. This is common in medical datasets, where normal tissue or background may dominate while rare pathological structures occupy only a small region.

Early stopping is the practice of interrupting training when validation performance stops improving, in order to reduce overfitting.

2.4 Architectures for images and sequences

The architectural terminology in this subsection follows the foundational literature on convolutional, recurrent, and biomedical segmentation networks [6], [12], [34], [35], [36].

Convolutional neural networks (CNNs) are neural networks designed especially for grid-structured data such as images. They are strong at learning local visual patterns such as edges, textures, and shapes.

Convolution is the operation in which a small learnable filter moves across an image and computes a local response at each position. This is how CNNs detect recurring patterns.

A filter (or kernel) is the small learnable pattern used in a convolution. Different

filters learn to respond to different kinds of structures.

Pooling is an operation that reduces the spatial size of a representation, usually to save computation and to retain the most important information.

The receptive field of a neuron or feature is the region of the input image that can influence it. A larger receptive field means that the model can use broader context.

Locality bias is the natural tendency of CNNs to focus first on nearby information. This is often helpful, but it can make it harder to capture very long-range relationships across an entire image.

An encoder is the part of a model that transforms the input into a more abstract internal representation.

A decoder is the part of a model that transforms an internal representation back into an output form, such as a segmentation map.

An encoder-decoder architecture is a model design that first compresses or summarizes information and then reconstructs a task-specific output from that summary.

A bottleneck is the most compressed or information-dense part of an encoder-decoder model. It often contains high-level information but less spatial detail.

A skip connection is a direct path that passes information from an earlier layer to a later one. In segmentation, skip connections help preserve fine spatial detail.

A residual connection is a special kind of skip connection in which a layer learns a refinement added to its input. Residual connections often make deep models easier to train.

U-Net is a well-known encoder-decoder architecture designed for biomedical image segmentation. Its characteristic feature is the use of skip connections between encoder and decoder stages.

Normalization is a family of techniques used to keep internal activations numerically well-behaved during training. Common forms include Batch Normalization and Layer Normalization.

Recurrent neural networks (RNNs) are models designed for sequences, such as time series, text, or ordered signals. They process information step by step and maintain an internal memory of previous elements.

Long short-term memory (LSTM) and **gated recurrent unit (GRU)** are improved forms of RNNs designed to remember useful information over longer ranges than a simple recurrent network.

2.5 Attention and transformer terminology

The definitions in this subsection follow the original attention and Transformer literature together with later vision-specific adaptations [7], [8], [37].

The attention mechanism is a method that allows a model to decide which parts of the input are most relevant for the current computation. It is a way of assigning importance rather than treating every part of the input equally.

A transformer is a neural architecture built mainly around attention mechanisms rather than recurrence or only convolution. It is powerful because it can model long-range relationships very effectively.

Self-attention is the form of attention in which different parts of the same input attend to one another. In an image model, one patch can therefore directly use information from many other patches.

Query, key, and value vectors are the three learned representations used inside attention. A query asks what information is needed, a key describes what information is available at a location, and a value contains the information that can be passed forward if the query and key match well.

Multi-head attention means that the model performs several attention operations in parallel. Each head can learn a different type of relationship, such as shape, boundary, or global context.

A token is one discrete unit processed by a Transformer. In language, a token may be a word or word fragment. In vision, a token usually corresponds to an image patch or another learned region representation.

A patch is a small block cut from an image. Vision Transformers often divide an image into patches and treat each patch as one token.

Patch embedding is the process of converting each image patch into a numerical vector that the Transformer can process.

Positional encoding is additional information that tells the Transformer where each token comes from. Without positional information, the Transformer would know what the pieces are but not how they are arranged.

Rotary positional embeddings are a particular way of encoding position by rotating internal vector components according to location. Their purpose is to preserve relative positional relationships in a flexible way.

A feed-forward network (FFN) inside a Transformer is a small neural subnetwork applied after attention. It refines the token representations independently at each position.

A vision transformer (ViT) is a Transformer adapted for image analysis. Instead of reading words, it reads image patches.

A hybrid CNN-transformer model combines convolutional layers and Trans-

former layers in one system. The usual motivation is to benefit from both local detail extraction and global context modeling.

2.6 Medical imaging and biomedical observation

The medical imaging terminology in this subsection is consistent with major surveys on deep learning in biomedical imaging and medical segmentation [2], [3].

A pixel is the smallest two-dimensional element of a digital image.

A voxel is the three-dimensional counterpart of a pixel. It is used in volumetric imaging, such as MRI or CT scans.

Image classification is the task of assigning one label to an entire image.

Object detection is the task of finding where objects are located in an image and assigning a class to each one, usually by drawing bounding boxes.

Semantic segmentation is the task of assigning a class label to every pixel in the image. It tells us what each location represents.

A segmentation mask is the image-sized map of labels used to represent a segmentation result or a segmentation annotation.

Instance segmentation goes one step further than semantic segmentation by separating individual objects of the same class from one another. For example, it does not only mark "cell" but distinguishes one cell from the next.

Magnetic resonance imaging (MRI) is a medical imaging technique that uses magnetic fields and radio waves to produce detailed images of internal body structures, especially soft tissues.

Computed tomography (CT) is a medical imaging technique that reconstructs cross-sectional images of the body from many X-ray measurements taken at different angles.

Microscopy is the use of specialized instruments to observe structures too small to be seen clearly with the unaided human eye.

Histopathology is the study of diseased tissue under the microscope. In AI research, histopathology images are often used for classification, detection, and segmentation tasks.

Hematology is the branch of medicine and biology concerned with blood, blood-forming organs, and blood-related disorders.

A blood smear is a thin layer of blood spread on a glass slide for microscopic examination. It is one of the key data sources in this thesis.

Morphology refers to the form, shape, and visible structure of a biological object. In blood analysis, morphology includes properties such as size, contour, texture, and internal appearance.

Staining is the process of applying chemical dyes to biological material so that structures become easier to see under the microscope.

A staining artifact is an unwanted visual effect introduced by sample preparation, staining, imaging, or handling. Artifacts can confuse both human experts and machine learning systems.

Inter-observer variability means that different experts may provide slightly different interpretations or annotations for the same image. This is a major source of uncertainty in medical datasets.

2.7 Fuzzy logic, uncertainty, and ambiguous boundaries

The concepts in this subsection are drawn from the standard theory of fuzzy sets, fuzzy inference, fuzzy clustering, and type-2 fuzzy logic [17], [18], [19], [20].

Boolean or crisp logic is the classical system in which a statement is either completely true or completely false. Many real biomedical phenomena do not fit this rigid structure well.

Fuzzy logic is a framework in which truth can take intermediate values between zero and one. It is useful when boundaries are gradual, uncertain, or partly overlapping.

A fuzzy set is a set in which an element can belong partially rather than only fully belonging or fully not belonging.

A membership degree is the number that expresses how strongly an element belongs to a fuzzy set. A value close to one means strong membership, while a value close to zero means weak membership.

A membership function is the rule used to assign membership degrees.

A fuzzy rule is an interpretable rule written in a form such as "if a region is bright and round, then it is likely to belong to a certain class." Such rules make fuzzy systems intuitive and expressive.

A fuzzy inference system is a full reasoning framework that combines fuzzy sets, membership functions, and fuzzy rules to produce a decision or score.

Fuzzy clustering is clustering in which a data point can belong to several groups at the same time with different membership degrees.

Fuzzy C-means (FCM) is a widely used fuzzy clustering method. Instead of assigning each data point to exactly one cluster, it learns soft memberships across multiple clusters.

Type-2 fuzzy logic extends ordinary fuzzy logic by allowing uncertainty not only in the data but also in the membership values themselves. It is useful when the uncertainty is especially strong or difficult to model precisely.

Ambiguity means that the available evidence supports more than one plausible

interpretation. In medical images, ambiguity often appears at borders, in low-contrast regions, or in overlapping structures.

Uncertainty is the broader concept that captures the model's lack of complete certainty about a prediction. A good medical model should not only predict but also reflect when the evidence is weak.

A pseudo-label is a label generated automatically by a model for previously unlabeled data. Pseudo-labels are commonly used in semi-supervised learning, but they must be treated carefully because they may contain errors.

2.8 Evaluation, reliability, and experimental methodology

The evaluation vocabulary in this subsection follows the standard methodology of machine learning experiments and medical image segmentation studies [2], [3].

A baseline is a simple or established reference method used for comparison. New methods are meaningful only when they are compared against credible baselines.

A benchmark is a standardized comparison setting used to evaluate multiple methods on the same task.

An ablation study is a controlled experiment in which one component of a method is removed or changed so that its specific contribution can be measured.

Cross-validation is an evaluation strategy in which the dataset is split into several parts and the experiment is repeated multiple times with different train-test partitions. It provides a more robust estimate of performance when data are limited.

A confusion matrix is a table that summarizes how often predicted classes match the true classes and where the errors occur.

Accuracy is the proportion of predictions that are correct. It is easy to understand, but it can be misleading when one class is much more common than the others.

Precision measures how many items predicted as positive are actually positive. High precision means few false alarms.

Recall measures how many of the truly positive items were successfully found. In medicine, recall is often very important because missed abnormalities can be costly.

Sensitivity is another name for recall in many medical contexts.

Specificity measures how well the system correctly identifies negative cases as negative.

The F1-score is the harmonic mean of precision and recall. It is useful when both false positives and false negatives matter.

Intersection over union (IoU) measures the overlap between a predicted region and the true region. It is calculated as the intersection divided by the union of the two regions.

Mean intersection over union (mIoU) is the average IoU across classes.

Dice similarity coefficient (Dice or DSC) is another overlap measure commonly used in medical segmentation. It is closely related to IoU and is especially popular in biomedical literature.

Pixel accuracy measures the fraction of pixels that are assigned the correct class.

A confidence score is the numerical strength with which a model supports a prediction.

Calibration describes how well predicted confidence matches actual correctness. A well-calibrated model gives high confidence when it is usually right and lower confidence when it is often wrong.

Interpretability refers to methods that help a human understand why the model produced a certain result. This is important in scientific and medical settings because users must be able to assess whether the model is reasoning in a sensible way.

2.9 Signal processing and efficient computation

The terminology in this subsection follows standard signal-processing and efficient deep learning practice, especially in the context of large vision models [8], [27], [38].

The spatial domain is the ordinary image view in which information is represented directly by position, such as rows and columns of pixels.

The frequency domain is a different way of describing a signal, focusing on how quickly or slowly patterns change rather than where they are located. Smooth trends correspond to low frequencies, while sharp edges and fine detail correspond to high frequencies.

The Fourier transform is the mathematical tool that converts a signal from the spatial domain into the frequency domain.

The fast Fourier transform (FFT) is an efficient algorithm for computing the Fourier transform. It is important in this thesis because it offers a practical way to process large image representations more efficiently.

Computational complexity describes how the amount of computation grows as the problem becomes larger.

Quadratic complexity, written as $\mathcal{O}(N^2)$, means that the computation grows roughly with the square of the input size. This can quickly become expensive for high-resolution images with many tokens.

A graphics processing unit (GPU) is specialized hardware designed to perform many numerical operations in parallel. Modern deep learning relies heavily on GPUs because neural network training is computationally demanding.

Memory footprint is the amount of computer memory required to store the model, its intermediate activations, and related data during training or inference.

Mixed precision is a strategy in which some calculations are performed with lower numerical precision in order to save memory and increase speed while preserving sufficient accuracy.

FP32 means 32-bit floating-point arithmetic, while **FP16** means 16-bit floating-point arithmetic. FP16 is smaller and faster, but it can be more numerically delicate.

Out-of-Memory (OOM) is the failure that occurs when the available memory is insufficient for the requested computation. OOM is a practical limitation in high-resolution Transformer-based image analysis.

Having established the terminology, I now move from definitions to formal theory. The next section provides the mathematical and algorithmic foundations that I use throughout the experimental chapters. It should therefore be read as the conceptual bridge between the introductory material and the empirical sequence that follows.

3 Theoretical foundations of artificial intelligence

Before presenting the experimental studies, I summarize only the theoretical ideas that are required later in the thesis. The focus is on the mathematical description of trainable models, dense image analysis, and Transformer-based segmentation.

3.1 Mathematical and computational preliminaries

The thesis relies mainly on linear algebra, multivariate calculus, and probability. Images, feature maps, and parameter sets are represented as tensors. The basic operation of a neural layer is an affine mapping followed by a non-linear transformation,

$$\mathbf{z} = \mathbf{W}\mathbf{x} + \mathbf{b}, \tag{1}$$

and learning is formulated as minimization of a loss function $\mathcal{L}(\theta)$ with respect to model parameters θ. The basic gradient-descent update is

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla_\theta \mathcal{L}(\theta^{(t)}), \tag{2}$$

where η is the learning rate. This notation is sufficient for the optimization procedures, segmentation losses, and attention-based models used in the later chapters.

3.2 The artificial neuron and activation functions

The artificial neuron is the basic computational unit of the models used in this dissertation. For an input vector $\mathbf{x}$, its output is

$$a = f(\mathbf{w}^T\mathbf{x} + b), \tag{3}$$

where $\mathbf{w}$ are learnable weights, b is a bias term, and $f(\cdot)$ is a non-linear activation function. Non-linearity is essential, because without it a deep network would collapse into a single linear transformation. In modern vision models, ReLU and its variants dominate because they are simple, computationally efficient, and effective in deep optimization.

3.3 Artificial neural networks (ANNs) and deep learning

An Artificial Neural Network is a layered composition of such neurons. In feedforward form, the output of each layer becomes the input of the next layer, so the full model can be written as a nested function composition. Deep learning refers to the case in which several hidden layers are used to learn increasingly abstract representations directly from data. In this thesis, that hierarchy is important because early layers

capture local visual patterns, whereas deeper layers encode more task-specific semantic structure.

3.4 The backpropagation algorithm

Training requires an efficient way to compute how each parameter affects the loss. Backpropagation does this by applying the chain rule from the output layer back through the network. For a parameter θ, the update direction is determined by $\partial \mathcal{L}/\partial \theta$. In practice, backpropagation is the mechanism that makes end-to-end optimization of CNNs, Transformers, and segmentation networks feasible.

3.5 Convolutional neural networks (CNNs)

Convolutional Neural Networks remain one of the core building blocks of the thesis. A convolution applies a learned filter locally across the image,

$$(\mathbf{X} * \mathbf{K})(i,j) = \sum_m \sum_n \mathbf{X}(i+m, j+n)\, \mathbf{K}(m,n), \tag{4}$$

producing feature maps that respond to edges, textures, shapes, and progressively more complex structures. Their main advantages are locality, weight sharing, and translation-sensitive feature extraction. For segmentation, CNNs are often organized as encoder-decoder systems with skip connections, as in U-Net, so that coarse semantic information and fine spatial detail can be combined. These ideas remain important even in the later Transformer-based models, whose decoders still rely on spatial reconstruction principles developed in CNN-based segmentation.

3.6 The attention mechanism

Attention allows a model to weight different parts of an input according to their relevance. In the scaled dot-product form used later in the thesis,

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \tag{5}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are the query, key, and value matrices. This mechanism is crucial because it enables direct modeling of long-range dependencies, which standard convolutions capture only indirectly through depth and receptive-field growth. Its main limitation is computational cost, since dense self-attention scales poorly with the number of tokens.

3.7 The transformer architecture

A Transformer stacks self-attention blocks with position-wise feed-forward networks, residual connections, and normalization layers. In encoder form, each layer updates a sequence representation by combining global contextual aggregation with local nonlinear transformation. For the present thesis, the key property of Transformers is their ability to relate distant image regions directly. That property is central to microscopy segmentation, where weak boundaries and scattered structures often require context beyond a local convolutional window.

3.8 Vision transformers (ViT)

Vision Transformers adapt the Transformer to images by splitting an image into patches, projecting them to token embeddings, and processing the resulting sequence with self-attention layers. This replaces the strictly local inductive bias of CNNs with a more global representation mechanism. In practice, modern vision models often modify the original ViT idea through hierarchical designs, patch merging, and segmentation-oriented decoders. Those developments are directly relevant to this dissertation, because the later TSSN and the final semi-supervised hematological framework both rely on Transformer principles while still preserving the spatial detail required for dense prediction.

3.9 Advanced optimization and regularization

The models used in the thesis are trained with adaptive optimizers, most often Adam or Adam-like variants, together with regularization strategies such as weight decay, augmentation, dropout, and normalization. Adam updates parameters using running estimates of the first and second moments of the gradient,

$$\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\mathbf{g}_t, \tag{6}$$

$$\mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1-\beta_2)\mathbf{g}_t^2, \tag{7}$$

$$\theta_t = \theta_{t-1} - \eta \frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t} + \epsilon}. \tag{8}$$

These techniques are important in the later chapters because the datasets are limited, the class distributions are imbalanced, and the final segmentation systems must remain stable despite architectural complexity and annotation ambiguity.

4 Malaria detection using deep learning

The first experiment, based on initial works from [39], [40], is close to the final target domain of this thesis. The malaria study addresses blood-smear imagery and dense cellular analysis directly. It therefore functions as the immediate precursor to the final framework, allowing me to refine segmentation, preprocessing, and system-integration strategies within a hematological setting before combining them with Transformers, semi-supervision, and fuzzy logic.

4.1 Introduction and comprehensive background

Malaria remains one of the most formidable and persistent global health challenges of our era, a life-threatening disease orchestrated by protozoan parasites of the genus *Plasmodium*. These parasites are transmitted to hosts through the bite of infected female *Anopheles* mosquitoes. The disease's impact is profound, with the World Health Organization reporting an estimated 282 million clinical cases and 610,000 deaths worldwide in 2024, a burden disproportionately borne by low-income and developing regions, particularly in sub-Saharan Africa [41]. The clinical and economic ramifications of malaria are staggering, perpetuating cycles of poverty and straining fragile healthcare infrastructures. The imperative for rapid, accurate, and accessible diagnostic methodologies is therefore not merely a scientific objective but a moral and pragmatic necessity in the global public health arsenal.

Traditionally, the confirmatory diagnosis of malaria has relied upon the microscopic examination of Giemsa-stained thick and thin blood smears, a technique regarded as the historical gold standard. This method, while specific, is fraught with significant limitations. It demands extensive expertise from highly trained microscopists, is inherently time-consuming and labor-intensive, and its accuracy is susceptible to human fatigue, inter-observer variability, and the subjective interpretation of subtle morphological features [42]. These challenges are acutely felt in resource-limited settings and rural clinics where specialist personnel are scarce. Furthermore, the expanding geographical range of malaria vectors and parasites, fueled in part by climate change, globalization, and increased travel, introduces the disease to previously non-endemic regions and naive populations, including both humans and animals. This dynamic landscape underscores the critical need for diagnostic tools that are not only accurate but also scalable, rapid, and deployable in diverse field conditions.

In parallel, the field of veterinary medicine confronts a similar, though often underrecognized, threat from malarial parasites. A diverse array of *Plasmodium* species infects a wide taxonomic range of animals, including birds, non-human primates, ungulates, and bats. Avian malaria, caused by parasites like *Plasmodium*

relictum, has been implicated in devastating population declines and extinctions of endemic bird species in ecosystems such as Hawaii [43]. In mammals, from chimpanzees harboring *P. reichenowi* to white-tailed deer infected with *P. odocoilei*, these parasites form complex enzootic cycles. The zoonotic potential of certain species, most notably *P. knowlesi* from macaques, highlights the porous boundary between animal reservoirs and human health [44]. Diagnosing malaria in animals faces identical technical hurdles as in humans-reliance on skilled microscopy for blood smear analysis-but is further hampered by even greater constraints in veterinary diagnostic resources and expertise.

This confluence of challenges in both human and veterinary medicine creates a fertile ground for technological innovation. The advent of sophisticated artificial intelligence (AI), particularly deep learning, presents a paradigm-shifting opportunity [1], [2]. Convolutional Neural Networks (CNNs), inspired by the biological visual cortex, have demonstrated exceptional proficiency in automating image analysis tasks, including classification, object detection, and segmentation. Within medical imaging, CNNs have achieved diagnostic performance comparable to experts in fields ranging from dermatology to radiology [2], [4]. For malaria diagnosis, the application of AI promises to transcend the limitations of manual microscopy by offering objective, rapid, and high-throughput analysis of digital blood smear images [39].

Previous research efforts have explored various deep learning architectures for malaria detection, including standard CNNs for image-level classification, object detection models like YOLO for bounding box localization, and segmentation-oriented frameworks [39], [45]. While these approaches have shown promising results, they often grapple with trade-offs between accuracy, computational complexity, and granularity of output. Classification networks provide a binary label but no spatial information; bounding box methods offer localization but with coarse, rectangular approximations that poorly capture the irregular, overlapping morphology of blood cells; and precise segmentation can be computationally demanding for real-time use on constrained hardware.

In my research, I address these gaps by proposing and developing a novel, custom-designed semantic segmentation convolutional neural network architecture specifically optimized for the task of malaria parasite detection in blood smear imagery. Semantic segmentation provides a per-pixel classification of the input image, assigning each pixel to a specific category (e.g., background, healthy cell, infected cell). This approach yields a highly detailed segmentation mask that precisely delineates the contours of each cell, a level of granularity that is superior to bounding boxes and provides richer diagnostic information than a simple image-level label. My architecture is engineered to balance high accuracy with computational efficiency, ensuring its potential viability for deployment on cost-effective hardware in field

settings.

The core intellectual contribution of this work is a holistic, cross-species framework for malaria diagnosis. I have designed a unified deep learning solution trained on a composite dataset of human blood smears, with the inherent flexibility to be adapted for veterinary use. The system functions as a sophisticated assistive tool: it processes a digital image from a standardized light microscopy setup, generates a precise semantic segmentation mask highlighting suspected infected cells, and presents this mask alongside the original image for rapid validation by a human specialist (medical or veterinary). This semi-automated workflow aims to drastically reduce diagnostic time, mitigate observer fatigue, minimize misclassification rates, and enhance overall detection sensitivity, particularly for low-level parasitemia. By augmenting, rather than replacing, the expertise of the diagnostician, the system leverages the strengths of both artificial and human intelligence.

4.2 Epidemiology, parasitology, and clinical significance

To fully appreciate the context and necessity of my diagnostic innovation, a deep understanding of malaria's biological and clinical landscape is essential. Malaria is caused by apicomplexan parasites of the genus *Plasmodium*. While over 150 species exist, only a handful routinely infect humans: *P. falciparum*, *P. vivax*, *P. malariae*, *P. ovale*, and the zoonotic *P. knowlesi*. *P. falciparum* is the most virulent, responsible for the vast majority of malaria-related mortality, particularly in sub-Saharan Africa. Its pathogenicity is linked to unique mechanisms like cytoadherence, which leads to sequestration of infected erythrocytes in deep vascular beds, causing severe complications such as cerebral malaria. *P. vivax*, while less frequently fatal, has the widest geographical distribution and possesses a latent liver stage (hypnozoite) that can cause relapses months or years after the initial infection, complicating treatment and eradication efforts. *P. knowlesi*, naturally infecting long-tailed and pig-tailed macaques in Southeast Asia, presents a significant zoonotic threat with a rapid, potentially fatal clinical course resembling falciparum malaria.

The parasite's life cycle is a complex, heteroxenous process involving both a mosquito vector and a vertebrate host. When an infected female *Anopheles* mosquito takes a blood meal, it inoculates sporozoites from its salivary glands into the host's dermis. These motile sporozoites rapidly enter the bloodstream and migrate to the liver, where they invade hepatocytes. This initiates the pre-erythrocytic (or exo-erythrocytic) stage, an asymptomatic period of 7-14 days where the parasite undergoes asexual replication (schizogony), culminating in the rupture of the liver schizont and the release of thousands of merozoites into the bloodstream. For *P. vivax* and *P. ovale*, a subset of sporozoites may develop into dormant hypnozoites

within hepatocytes, establishing a reservoir for future relapse.

The released merozoites commence the erythrocytic stage, the phase responsible for all clinical manifestations of malaria. Merozoites invade red blood cells (RBCs), where they progress through ring, trophozoite, and schizont morphological stages. The mature schizont eventually ruptures the RBC, releasing a new generation of merozoites to infect fresh erythrocytes. This cyclic, synchronous destruction of RBCs correlates with the classic periodic fevers (paroxysms) of malaria. Some merozoites within the bloodstream instead differentiate into sexual forms called gametocytes. When a mosquito ingests these gametocytes during a blood meal, they undergo sexual reproduction in the mosquito's midgut, leading to the formation of an oocyst. Sporozoites develop within the oocyst, migrate to the salivary glands, and complete the transmission cycle.

The clinical presentation of malaria is notoriously variable, ranging from asymptomatic parasitemia to severe, life-threatening disease. Symptoms typically include fever, chills, headache, myalgia, and fatigue. Severe malaria, most commonly caused by *P. falciparum*, is a medical emergency characterized by complications such as cerebral malaria (impaired consciousness, seizures), severe malarial anemia, acute respiratory distress syndrome, metabolic acidosis, hypoglycemia, and multi-organ failure. The pathogenesis of severe disease is multifactorial, involving massive hemolysis, microvascular sequestration of parasitized RBCs, endothelial dysfunction, and a dysregulated, excessive host inflammatory response. Vulnerable populations include young children, pregnant women, non-immune travelers, and immunocompromised individuals.

In the veterinary context, the biological principles are analogous, but the cast of characters changes. Avian malaria parasites (e.g., *P. relictum*, *P. gallinaceum*) can cause devastating outbreaks in naive bird populations, as historically seen in Hawaii. In non-human primates, a plethora of *Plasmodium* species exist, some of which are phylogenetic sisters to human parasites, offering critical insights into malaria evolution. Other mammalian hosts, like deer infected with *P. odocoilei* or bats with their unique hepatocystis parasites, maintain enzootic cycles. The rising prominence of *P. knowlesi* as a human pathogen epitomizes the zoonotic bridge and underscores the importance of a One Health perspective that integrates human, animal, and environmental health.

Diagnosis remains the critical gateway to effective management. Beyond clinical suspicion, laboratory confirmation is mandatory. While rapid diagnostic tests (RDTs) that detect parasite antigens are valuable for quick screening, light microscopy of Giemsa-stained blood smears retains its status as the cornerstone of diagnosis where resources allow. The thick smear is used for sensitive detection of parasites, while the thin smear allows for species identification and staging based on parasite mor-

phology within RBCs. However, as previously emphasized, this method's accuracy is inextricably linked to the skill and endurance of the microscopist. Molecular techniques like polymerase chain reaction (PCR) offer superior sensitivity and specificity for species confirmation and detecting low-level infections but are costly, require sophisticated laboratory infrastructure, and have longer turnaround times, rendering them impractical for point-of-care use in most endemic settings.

It is against this backdrop of a pervasive disease with a demanding, expertise-reliant diagnostic standard that my research intervenes. The automation of blood smear analysis through deep learning directly attacks the bottlenecks of time, cost, and human resource limitations. By providing a consistent, rapid, and accurate first-pass analysis, an AI system can democratize access to high-quality malaria diagnostics, empower healthcare and veterinary workers, and ultimately contribute to timely treatment and improved patient and animal outcomes.

4.3 Foundations of deep learning and semantic segmentation

The general concepts of CNNs, semantic segmentation, encoder-decoder models, U-Net, adaptive optimization, and augmentation were introduced earlier in Section 2 and Sections 3.5–3.9.

For malaria detection, the central methodological decision is to treat the task as pixel-wise segmentation rather than image classification or coarse object detection. Parasites occupy small, irregular, and often overlapping regions, so the clinically relevant output is a detailed mask rather than a single image label or a bounding box. For that reason, this study adopts a U-Net-inspired encoder-decoder CNN with skip connections, trains it with NAdam, and relies on strong augmentation to compensate for limited annotations and staining variability.

4.4 Dataset curation and advanced preprocessing pipeline

The performance of any deep learning model is fundamentally contingent on the quality, quantity, and relevance of the data on which it is trained. For this research, I utilized a publicly available benchmark dataset known as "Malaria Bounding Boxes", hosted on Kaggle.com. This composite dataset is particularly valuable as it amalgamates blood smear images from three distinct sources, thereby introducing beneficial variability: one set from Brazil (curated by researcher Stefanie Lopes), one from Southeast Asia (Benoit Malleret), and a time-course series (Gabriel Rangel). In total, the dataset comprises 1,364 images, containing annotations for approximately 80,000 individual cells.

Blood samples were prepared according to standard clinical hematology protocols: venous blood was collected, thin blood films were made on glass slides, and stained

using Giemsa reagent, which differentially stains cell nuclei (purple) and cytoplasm (pink). This results in the characteristic appearance where malaria parasites within red blood cells are visible as purple chromatin dots and/or pinkish cytoplasm.

The original annotation schema provided with the dataset was detailed, reflecting the biological complexity of the infection. Annotators, under the supervision of malaria expert Stefanie Lopes, labeled each cell with a bounding box and assigned it to one of six possible classes:

- **Uninfected cells:** Red Blood Cells (RBCs) and Leukocytes (White Blood Cells).
- **Infected cells:** Four developmental stages of the parasite within RBCs-Gametocytes, Rings, Trophozoites, and Schizonts.
- **Difficult cells:** Cells where the annotator was uncertain about the class assignment.

A prominent characteristic of this dataset, mirroring biological reality, is a severe class imbalance. Uninfected RBCs constitute over 95% of all annotated objects, while infected cells and leukocytes are relatively rare. This imbalance poses a challenge for training, as a naive model might learn to simply predict "uninfected RBC" for every pixel and still achieve a high overall accuracy, while failing miserably at the critical task of detecting the rare positive cases.

To tailor the dataset for my specific research objective-binary segmentation of infected versus healthy cells-I implemented a streamlined labeling strategy. I consolidated the original six classes into two super-classes:

1. **Healthy cells:** This class encompasses all uninfected RBCs and leukocytes.
2. **Infected cells:** This class aggregates all four parasitic stages (Rings, Trophozoites, Schizonts, Gametocytes). Crucially, I also assigned cells originally marked as "difficult" to the infected class. This decision is clinically prudent: the system should err on the side of caution, flagging any ambiguous cell for expert human review rather than risk missing a true infection.

The provided bounding box annotations, while useful, were suboptimal for training a per-pixel segmentation network. Rectangular boxes are inherently imprecise for irregularly shaped cells, include considerable background pixels, and create ambiguous regions where boxes overlap. Therefore, a critical and novel component of my preprocessing pipeline was the conversion of these coarse bounding box labels into precise, pixel-accurate semantic segmentation masks.

I developed a custom algorithm to perform this conversion automatically and robustly. The process for each bounding box is as follows:

1. The rectangular region is extracted from the grayscale or color image.
2. Within this region, the algorithm performs adaptive clustering of pixels based

on intensity. Giemsa-stained parasites and cell nuclei are typically darker than the surrounding RBC cytoplasm or background.

3. Using a combination of pre-computed average intensity values for cell interiors and custom heuristics (e.g., expected size, circularity), the algorithm separates the cluster of dark pixels corresponding to the cell (and any internal parasite) from the lighter background pixels within the box.
4. A morphological post-processing step is applied for artifact removal, smoothing the resultant binary mask for the cell.
5. Finally, this binary mask is placed into the corresponding channel of a multi-channel label image based on the cell's class (healthy or infected). Background pixels outside any mask are assigned a default value.

This algorithm effectively transforms the dataset from one with imprecise rectangular annotations to one with high-quality, pixel-level ground truth masks, which are essential for training a segmentation network effectively. Figure 1 illustrates examples of original images and their corresponding generated segmentation masks.

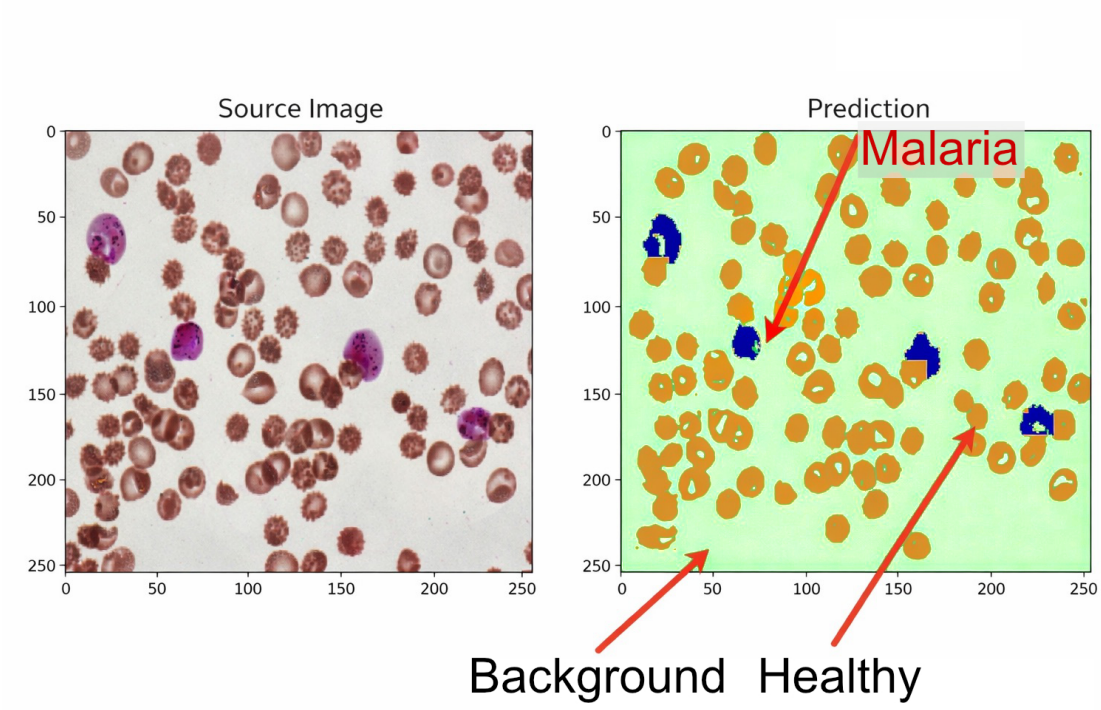


Figure 1: Illustration of the preprocessing pipeline. Left: Original blood smear image with bounding box annotations (green: healthy, red: infected). Right: The generated semantic segmentation mask after custom algorithm processing (Orange: Healthy Cells, Blue: Infected Cells, Green: Background).

Following mask generation, standard normalization was applied. All input images, originally 8-bit (values 0-255), were converted to 32-bit floating-point format and normalized by dividing each pixel value by 255.0, scaling the data to the range [0, 1].

This standardization stabilizes and accelerates the training process.

To combat the class imbalance and increase the effective size of the training set, I employed an extensive data augmentation regimen during training. For each batch, images and their corresponding masks were subjected to a random series of geometric and photometric transformations including: horizontal and vertical flipping, random rotation (within a +-30° range), random zoom (90%-110%), random shifts in hue and saturation, and the addition of small amounts of Gaussian noise. These augmentations simulate the natural variations encountered in different laboratories and staining batches, significantly improving the model's robustness and generalization capability. Empirical tests confirmed that this augmentation strategy boosted validation accuracy by over 7%.

4.5 Custom semantic segmentation neural network architecture

The cornerstone of my research is a deep convolutional neural network architecture designed specifically for the semantic segmentation of malaria parasites in blood smear images. The design philosophy prioritized a balance between predictive accuracy, computational efficiency, and a manageable parameter count for potential deployment on lower-power hardware. The model follows a symmetric encoder-decoder layout with residual blocks and U-Net-style skip connections.

Input and output specification: The network accepts a single RGB image of fixed dimensions as input. In the finalized model, I used a resolution of 300x300 pixels, which provides a good compromise between cellular detail and computational cost. The output is a 3-channel segmentation map of the same spatial dimensions (300x300). Each channel corresponds to a probability distribution over the pixels for one of the three abstract classes: Background, Healthy Cell, and Infected Cell. The final per-pixel class assignment is determined by taking the channel with the highest probability (argmax).

Encoder (contracting path): The encoder is responsible for feature extraction and contextual understanding. It consists of four sequential downsampling blocks. Each block performs the following operations:

1. **Two convolutional layers:** Each block uses 3x3 kernels, stride 1, and same padding. The number of filters doubles at each downsampling step, starting from 64 in the first block and increasing to 512 in the deepest block. Each convolution is followed by ReLU.
2. **Residual connection:** The input to the block is added to the output of the second convolutional layer, which stabilizes training and improves feature reuse.

3. **Max pooling:** A 2x2 max-pooling layer with stride 2 halves the spatial resolution and expands the effective receptive field.
4. **Dropout:** After pooling, a dropout layer with rate 0.3 regularizes the encoder.

Bottleneck: Following the fourth encoder block, the feature map is at its lowest spatial resolution (18x18 pixels given the 300x300 input) but its deepest channel dimension (512 filters). This bottleneck region captures the most abstract, high-level features of the entire image. My experiments indicated that a simple sequence of three convolutional layers (with 512 filters each, ReLU, and residual connections) here was sufficient. Adding more complexity to the bottleneck yielded diminishing returns in accuracy while significantly increasing training time.

Decoder (expansive path): The decoder's role is precise spatial localization and upsampling. It consists of four upsampling blocks, mirroring the encoder. Each decoder block performs:

1. **Transposed convolution:** A 2x2 transposed convolutional layer with stride 2 upsamples the feature maps.
2. **Feature concatenation:** The upsampled feature map is concatenated with the encoder feature map at the same resolution so that fine spatial detail is restored during decoding.
3. **Two convolutional layers:** Following concatenation, two 3x3 convolutional layers (with ReLU) process the combined features. The number of filters is halved at each step, moving from 512 down to 64.

Final layer: The output of the last decoder block passes through a final 1x1 convolutional layer with 3 filters, mapping each spatial location to logits for background, healthy cell, and infected cell. A channel-wise Softmax converts these logits into per-pixel class probabilities, yielding a 300x300x3 output tensor.

The total number of trainable parameters in this architecture is 95,354,474. While this is a substantial model, its efficiency is realized in the decoder's design and the use of bottlenecking, making evaluation (inference) relatively fast once trained.

Optimization and training details: I trained the network using NAdam, already introduced earlier as an Adam-family optimizer with Nesterov momentum. In this study I used $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and a constant learning rate of 0.00012, which provided stable convergence in empirical tuning. The loss function was per-pixel categorical cross-entropy between the Softmax probabilities and the one-hot encoded ground-truth mask. Training was conducted for 1000 epochs on a workstation equipped with an NVIDIA RTX 3090 GPU (24GB VRAM), an AMD Ryzen Threadripper 2950X CPU, and 128GB of RAM. The training curves in Figure 2 show stable convergence, with validation accuracy plateauing while the loss decreased steadily.

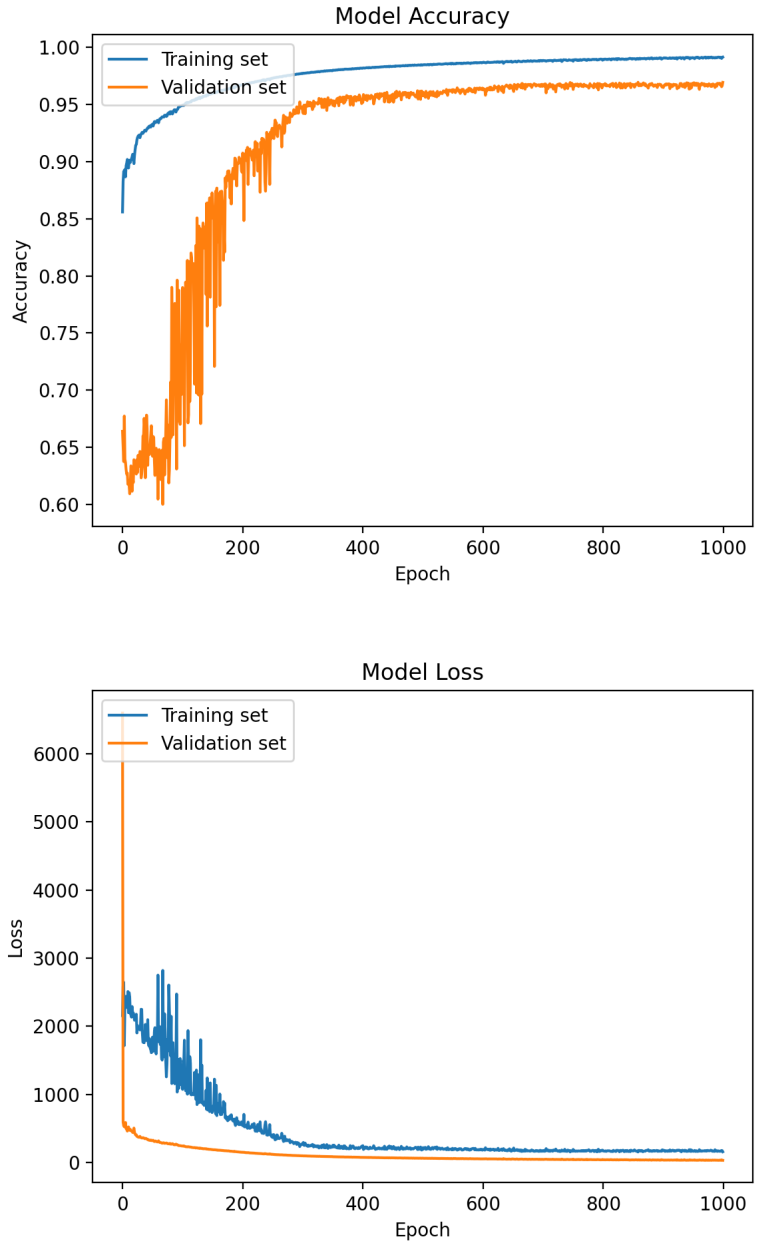


Figure 2: Model training and validation metrics over 1000 epochs. The top plot shows accuracy (training in blue, validation in orange) rising to over 99%. The bottom plot shows the corresponding loss (categorical cross-entropy) decreasing and stabilizing, indicating successful learning without severe overfitting.

4.6 Experimental results and performance evaluation

I conducted a rigorous evaluation of the proposed semantic segmentation model to quantify its diagnostic performance and compare it against contemporary state-of-the-art methods. The dataset of 1,364 images was split into training (70%), validation (15%), and test (15%) sets, ensuring no patient or sample overlap between sets.

Segmentation accuracy: The primary metric for the segmentation task is per-pixel accuracy-the percentage of pixels in the test set images that were correctly

classified into one of the three classes (Background, Healthy Cell, Infected Cell). My model achieved an impressive per-pixel accuracy of **97.1%** on the held-out test set. This high score indicates that the model is proficient at distinguishing fine cellular structures and boundaries from the background.

Clinical detection accuracy: In a clinical or veterinary setting, the precise contour of every single pixel is less critical than the correct identification and counting of infected versus healthy cells. Therefore, I implemented an auxiliary counting routine that processes the segmentation mask to generate bounding boxes around each connected component (cell cluster) and counts them. These counts are then compared to the ground truth cell counts derived from the original dataset's JSON annotations (after label simplification). Using this clinically relevant metric, the model's performance is even more remarkable, achieving an overall **99.68%** accuracy in correctly classifying cells as infected or healthy. This near-perfect score underscores the model's utility as a highly reliable screening tool.

Computational performance: Inference speed is crucial for practical deployment. On the high-end test hardware (NVIDIA RTX 3090), processing a single 300x300 image through the full pipeline (preprocessing, model inference, post-processing) takes approximately 22 milliseconds, equating to a throughput of over 45 frames per second. This demonstrates the feasibility of real-time analysis. The model's architecture, while deep, is sufficiently optimized to be ported to edge computing devices like the Google Coral TPU or NVIDIA Jetson, where inference would still take only a few seconds-well within acceptable limits for a point-of-care diagnostic.

Comparison with state-of-the-art: Table 1 places my model's performance in the context of other recent deep learning approaches for malaria detection published between 2019 and 2026. The comparison includes methods based on standard CNNs, YOLO variants, lightweight CNNs, Attention-based Deep CNNs (ADCN), Mask R-CNN, FPGA implementations, and more recent ensemble and explainable architectures. My semantic segmentation model (Seg-Sem CNN) achieves the highest reported accuracy at 99.68%, outperforming the next best (a Lightweight CNN at 98.85%) by a significant margin. This superior performance can be attributed to the precision of per-pixel segmentation, which avoids the localization inaccuracies inherent in bounding box methods and the information loss of simple image-level classification.

Qualitative visualization: The true strength of the semantic segmentation approach is vividly apparent in the visual output. Figure 3 showcases sample results from the test set. The left column displays the original blood smear image. The middle column shows the ground truth segmentation mask (for evaluation purposes). The right column presents the model's predicted color-coded overlay superimposed

Table 1: Comparative analysis of malaria detection models.

Article / model	**Type**	**Year**	**Accuracy**
My model (proposed) [39]	**Seg-Sem CNN**	**2022**	**99.68%**
Divyansh et al. [46]	CNN	2020	95.70%
Razin et al. [47]	YOLOv5 CNN	2022	96.21%
Alqudah et al. [48]	Lightweight CNN	2020	98.85%
Quan et al. [49]	ADCN	2020	97.47%
Turuk et al. [50]	Integrated CNN	2022	93.89%
Shekar et al. [51]	Fine-Tuned CNN	2020	95.99%
Rahman et al. [52]	TL-VGG16	2019	97.77%
Loh et al. [53]	Mask R-CNN	2021	94.57%
Saglam et al. [54]	FPGA CNN	2019	94.70%
Pongpanitanont et al. [55]	Ensemble Parallel NN	2026	97.00%
Parveen et al. [56]	Inception-ResNetV2	2025	97.80%
Shahin et al. [57]	Feature-fusion ensemble	2025	96.47%

on the original image.

The visualizations reveal several key strengths:

- **Precise delineation:** The model accurately traces the often irregular boundaries of both healthy and infected RBCs, even when cells are touching or slightly overlapping.
- **Accurate classification:** Infected cells, exhibiting the characteristic shape distortion and internal staining of parasites (visible as darker regions), are consistently highlighted in blue. Healthy cells, with their smooth, biconcave appearance, are marked in orange.
- **Robustness to Background:** The model effectively ignores background artifacts, stain debris, and white blood cells (which are part of the "healthy" class in my schema), focusing attention on the diagnostically relevant RBCs.
- **Intuitive output:** The final overlay provides an immediate, intuitive visual guide for the human expert. It directs their attention to areas of concern, significantly speeding up the verification process and reducing the cognitive load associated with scanning the entire smear manually.

Minor inaccuracies, when they occur, are typically confined to the very border pixels between a cell and the background, or in extremely rare cases where a heavily degenerated cell or a large clump of platelets might be misclassified. These errors are inconsequential for the clinical decision of "infected vs. not infected," as the cell center and overall morphology are correctly identified.

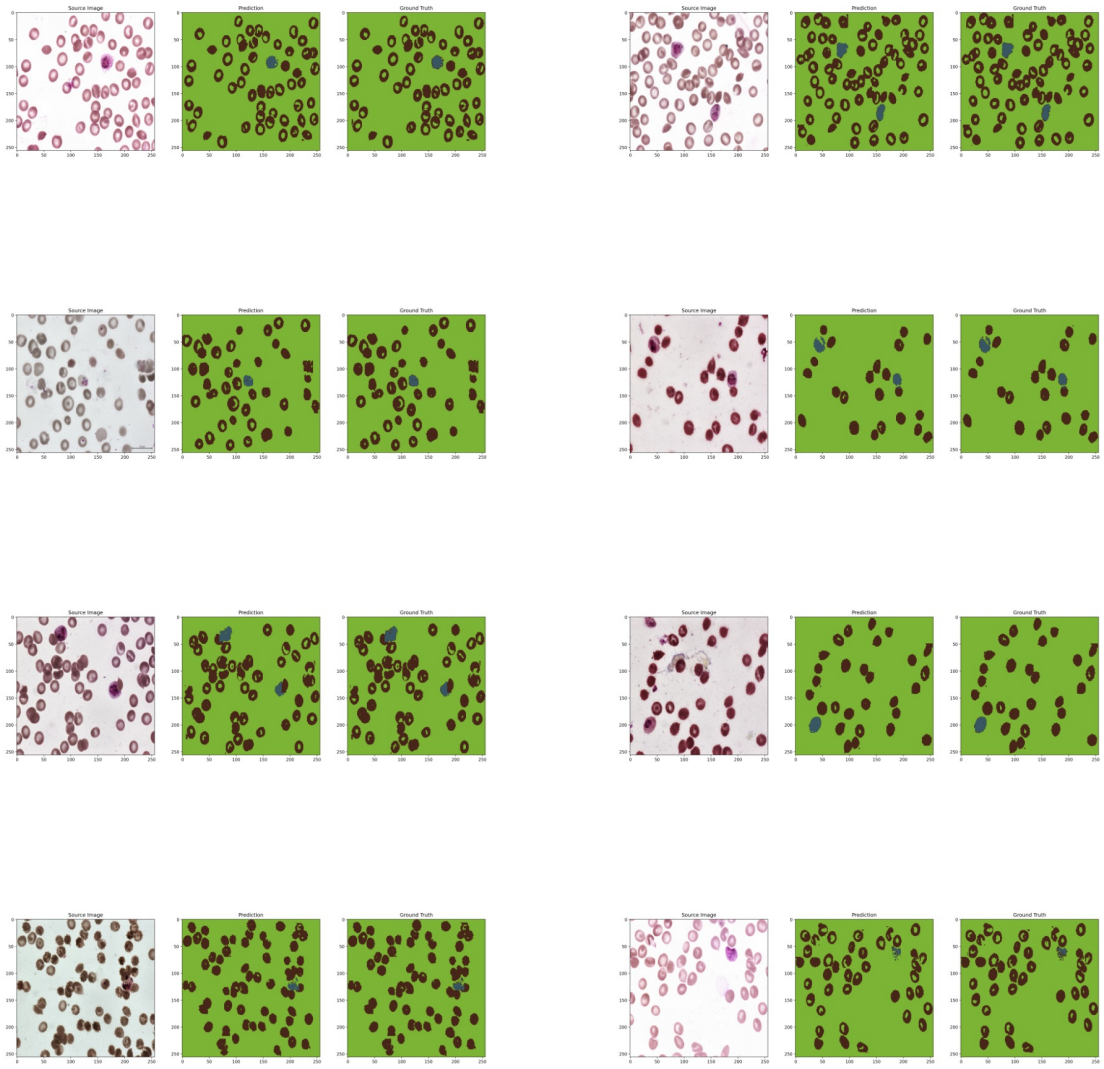


Figure 3: Qualitative results of the proposed system. Each row represents a different test sample. From left to right: (1) Original microscopic image, (2) Ground truth segmentation mask (Green=Background, Orange=Healthy, Blue=Infected), (3) Model prediction overlaid on the original image, providing an immediate diagnostic aid.

4.7 Discussion, limitations, and future research trajectories

The results presented herein affirm the viability and superior performance of a custom semantic segmentation neural network for the detection of malaria parasites in blood smear imagery. The achieved accuracy of 99.68% in cell-level classification establishes a new benchmark for automated malaria diagnosis. This success can be deconstructed into several key factors: the strategic choice of semantic segmentation over other paradigms, the meticulous design of a deep yet efficient encoder-decoder architecture with residual and skip connections, the development of a robust preprocessing pipeline to generate high-quality mask labels from bounding boxes, and the effective use of data augmentation to enhance generalization.

The proposed system offers transformative advantages over conventional mi-

croscopy:

- **Objectivity and Consistency:** The AI model provides a standardized, repeatable analysis, free from the subjective biases, fatigue, or skill variations that affect human microscopists.
- **Speed and Throughput:** Analysis that could take a human 10-30 minutes per smear is reduced to seconds, enabling high-volume screening in clinics or during epidemiological surveys.
- **Skill augmentation:** It empowers technicians and healthcare workers in peripheral settings by providing expert-level decision support, bridging the gap created by a shortage of specialist parasitologists.
- **Quantitative precision:** It delivers not just a qualitative "positive/negative" call but quantifiable data like parasite density, which is essential for treatment stratification and monitoring.
- **Cross-species potential:** While trained on human data, the fundamental approach of segmenting parasitic inclusions within erythrocytes is directly transferable to veterinary diagnostics with appropriate retraining on animal blood smear datasets.

Limitations:

1. **Class simplification:** Reducing the six original classes to two loses clinically valuable information. Distinguishing *P. falciparum* from *P. vivax* or identifying the parasite's life cycle stage (ring vs. schizont) can inform treatment choices and prognosis. My current model cannot provide this granularity.
2. **Artifact sensitivity:** While robust to many artifacts, the model could potentially be confused by extreme staining inconsistencies, excessive background debris, or certain blood disorders that dramatically alter RBC morphology (e.g., sickle cell anemia).
3. **Hardware dependency:** Real-time inference on truly low-cost, battery-powered edge devices (e.g., based on microcontrollers) remains a challenge. Further model compression techniques (pruning, quantization, knowledge distillation) would be required.

4.8 Conclusion

In this research work, I have presented a novel, end-to-end framework for the automated detection of malaria parasites in blood smear images, with inherent applicability to both human and veterinary medicine. At its core is a custom-designed semantic segmentation convolutional neural network that achieves a state-of-the-art cell classification accuracy of 99.68%. This performance stems from a meticulously engineered architecture combining an encoder-decoder structure with residual layers

and U-Net-inspired skip connections, optimized using the NAdam algorithm and trained on a carefully preprocessed dataset.

The implications of this research are significant. By democratizing access to high-quality diagnostic support, this technology has the potential to improve early detection rates, guide appropriate treatment, reduce mortality, and strengthen surveillance efforts in the global fight against malaria. While challenges remain in model generalization, regulatory approval, and hardware deployment, the foundational work presented here provides a robust and promising pathway forward.

5 Transformer-based semantic segmentation for medical imaging

The theoretical foundations developed so far establish the importance of custom design, computational efficiency, and transfer-based adaptation. The next empirical step is therefore to move directly to dense prediction to build up the necessary knowledge and experience needed for the final solution. In this chapter, I investigate Transformer-based semantic segmentation for medical imaging because pixel-wise delineation is the level of analysis ultimately required for the final hematological system. This experiment tests whether global self-attention can provide the global context that purely convolutional pipelines often capture only indirectly. This work has been initially started in [58].

5.1 Introduction

Medical imaging produces large amounts of visual data whose interpretation is important for diagnosis, treatment planning, and quantitative monitoring [1], [2], [3], [10]. In many settings, the clinically relevant question is not only whether an object is present, but also where it is located and how precisely its boundary can be delineated. For this reason, semantic segmentation has become a central task in computational medical imaging.

Convolutional Neural Networks (CNNs), especially encoder-decoder models such as U-Net, have been highly effective for biomedical segmentation [12]. Their strength lies in learning local texture and shape representations directly from data. However, their locality bias can make it difficult to incorporate long-range context when the interpretation of one region depends on distant structures elsewhere in the image.

Transformer architectures provide an alternative mechanism based on self-attention, through which each token can interact directly with all others [7], [8]. In vision, this property is attractive because it offers access to global context from early layers of the model. For medical imaging, where fine structures often appear in complex spatial environments, this can improve dense prediction if the computational cost is controlled appropriately [10], [14], [15].

In this chapter, I investigate a Transformer-based semantic segmentation architecture for dark-field microscopy. The specific task is the pixel-wise segmentation of spirochaete bacteria in microscopy images that also contain erythrocytes, debris, and imaging artifacts. My objective is to determine whether a Transformer-oriented encoder, combined with a lightweight decoder, can provide accurate segmentation while remaining suitable for detailed analysis of thin and visually challenging structures.

5.2 Related works

To properly contextualize the contributions of this research, it is necessary to survey the evolution of semantic segmentation in medical imaging, from classical image models to CNN-based dense prediction and finally to Transformer-based systems [10].

Pre-deep learning era: classical image processing. Before deep networks became dominant, medical segmentation was built on task-specific image models. Thresholding methods such as Otsu's criterion were attractive because they were simple and fast, active contours introduced deformable boundary fitting, level-set formulations improved topological flexibility, and atlas-based pipelines transferred labels from registered exemplars in anatomically structured domains [59], [60], [61], [62]. These approaches remain historically important, but the review literature consistently notes their sensitivity to initialization, noise, weak boundaries, and anatomy outside the assumptions of the chosen model [60], [61], [62].

The deep learning revolution: CNN dominance. The deep-learning transition changed the field by learning hierarchical features directly from data. FCN introduced dense prediction with fully convolutional inference, U-Net became the dominant biomedical architecture because skip connections preserved localization under limited data, V-Net extended the encoder-decoder idea to volumetric segmentation, and DeepLab-style atrous convolutions improved context capture without aggressive downsampling [11], [12], [13], [63]. Much of modern medical segmentation can be read as a sequence of refinements to this encoder-decoder template [10].

That refinement process is visible in several influential segmentation architectures. SegNet emphasized memory-efficient encoder-decoder inference by reusing pooling indices during upsampling [64]. Attention U-Net introduced gating to suppress irrelevant activations and focus the decoder on target anatomy [65]. UNet++ redesigned skip pathways to reduce the semantic gap between encoder and decoder representations [66]. DoubleU-Net stacked two encoder-decoder stages and used ASPP to improve refinement on difficult medical boundaries [67]. Roy et al. showed that spatial and channel squeeze-and-excitation blocks can recalibrate feature responses inside fully convolutional medical segmenters with consistent gains [68]. nnU-Net then demonstrated that a large fraction of state-of-the-art biomedical segmentation performance can be recovered through self-configuring pipelines rather than task-specific manual tuning [69].

The advent of attention and transformers in vision. Attention mechanisms emerged first as supplements to CNNs rather than full replacements. Non-local operations provided a direct way to model long-range dependencies inside vision backbones, addressing one of the central weaknesses of purely local convolutions

[70]. Vision Transformer then showed that patch-token processing with self-attention could compete with or surpass strong CNNs when sufficient pretraining was available, which opened the door to transformer-based dense prediction [7].

Transformers for medical image segmentation. Medical segmentation adopted these ideas along two main lines. Hybrid models such as TransUNet and UNETR keep convolutional or encoder-decoder structure while inserting transformers to improve global context modeling, whereas more segmentation-oriented transformer systems such as SETR and Swin-Unet explore pure or hierarchical token processing with explicit efficiency considerations [9], [14], [15], [71]. Recent hybrid proposals such as BEFUnet continue this trend by mixing CNN locality with transformer context rather than discarding either inductive bias [16]. Recent reviews conclude that transformers are particularly attractive when long-range spatial relations matter, but that data hunger, computational cost, and localization of very fine structures remain open problems in medicine [10].

Gap analysis and contribution positioning. This gap is especially visible in microscopy. Reviews of hematological and blood-smear analysis show that the field still leans heavily on CNN-based detectors and segmenters, largely because annotation is expensive and many targets are small, thin, overlapping, and visually ambiguous [22], [23], [24]. My chapter therefore addresses a part of the literature where transformer-based segmentation is promising in principle, but where task-specific architectural adaptation and data-efficient training are still needed in practice [10], [22].

5.3 Theoretical foundations: from self-attention to vision transformers

This section establishes the mathematical and conceptual groundwork necessary to understand the proposed architecture. I begin with the core mechanism of self-attention and build up to the Vision Transformer and its application to segmentation.

The self-attention mechanism. The fundamental operation in a Transformer is **scaled dot-product attention**. Given an input sequence of token embeddings $X \in \mathbb{R}^{N \times D}$ (N tokens, D dimensions each), it is projected into three matrices: Queries Q, Keys K, and Values V, all in $\mathbb{R}^{N \times D_k}$ (typically $D_k = D$):

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V \tag{9}$$

where $W^Q, W^K, W^V \in \mathbb{R}^{D \times D_k}$ are learnable weight matrices. The attention output

is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{D_k}}\right)V \tag{10}$$

The matrix $A = \text{softmax}\left(\frac{QK^T}{\sqrt{D_k}}\right)$ is the attention matrix of size $N \times N$. Element A_{ij} represents the normalized "affinity" or "attention weight" that token i pays to token j. The scaling factor $\sqrt{D_k}$ stabilizes gradients by preventing the dot products from growing too large in magnitude.

Multi-head attention (MHA) extends this by performing the operation h times in parallel with different learned projection matrices, concatenating the results, and projecting once more:

$$\text{MHA}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \tag{11}$$

$$\text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{12}$$

This allows the model to jointly attend to information from different representation subspaces.

The transformer encoder block. A standard Transformer encoder layer consists of two sub-layers with residual connections and layer normalization:

$$Z^{(l)} = \text{LayerNorm}\left(X^{(l-1)} + \text{MHA}\left(X^{(l-1)}\right)\right) \tag{13}$$

$$X^{(l)} = \text{LayerNorm}\left(Z^{(l)} + \text{FFN}\left(Z^{(l)}\right)\right) \tag{14}$$

where FFN is a Feed-Forward Network, typically two linear layers with a non-linearity (e.g., GELU) in between: $\text{FFN}(x) = \text{GELU}(xW_1 + b_1)W_2 + b_2$.

Vision transformer (ViT) formulation. To apply this to images, ViT reshapes a 2D image $I \in \mathbb{R}^{H\times W\times C}$ into a sequence of flattened 2D patches $I_p \in \mathbb{R}^{N\times(P^2\cdot C)}$, where (P, P) is the patch resolution, $N = HW/P^2$ is the number of patches. Each patch is linearly projected to a D-dimensional embedding using a trainable matrix $E \in \mathbb{R}^{(P^2C)\times D}$. A learnable [class] token $x_{\text{class}} \in \mathbb{R}^{1\times D}$ is prepended to the sequence for classification tasks. Learnable 1D positional embeddings $E_{\text{pos}} \in \mathbb{R}^{(N+1)\times D}$ are added to retain spatial information:

$$z_0 = [x_{\text{class}}; I_p^1 E; I_p^2 E; ...; I_p^N E] + E_{\text{pos}} \tag{15}$$

This sequence z_0 is then fed through L Transformer encoder layers. For classification, the output corresponding to the class token is passed to a classifier head.

Adapting ViT for semantic segmentation. ViT is a classification model. For dense prediction tasks like segmentation, several strategies exist:

- **Patch-level classification**: The simplest approach classifies each output patch embedding. This yields coarse, patch-level segmentation.
- **Decoder-based approaches**: Use the Transformer encoder as a feature extractor. The sequence of image patch embeddings (discarding the class token) is reshaped into a 2D feature map and passed through a CNN decoder (e.g., a U-Net-like decoder) to upsample to the original image resolution.
- **Pure Transformer decoders**: Some works propose Transformer-based decoders, where cross-attention is used between encoder outputs and learned query embeddings to produce segmentation masks.

My proposed TSSN adopts the decoder-based approach but makes critical modifications to the encoder and the interface between encoder and decoder.

5.4 Dataset and preprocessing

Dataset description and statistical analysis. The "Bacteria detection with dark-field microscopy" dataset, consists of 366 grayscale images (though often stored in 3-channel format) and their corresponding 3-class masks. A detailed statistical analysis reveals the inherent challenges:

- **Class imbalance**: The background class dominates, typically comprising over 95% of pixels. Spirochaetes, being thin filaments, occupy a minuscule fraction (often below 1%). Erythrocytes occupy an intermediate proportion. This severe imbalance can bias a model toward predicting the majority class if not addressed.
- **Object size and shape variability**: Spirochaetes vary in length, curvature, and coil tightness. Erythrocytes, while generally disc-shaped, can appear slightly elliptical and vary in size. Some cells may be partially outside the field of view.
- **Density and Occlusion**: Images range from sparse to densely packed with overlapping bacteria and cells. This presents a challenge for instance separation.
- **Image quality variations**: Variations in focus, illumination intensity, and contrast are present. Artifacts like dust, scratches, and out-of-focus bubbles are common.

To quantify this, I computed the following dataset statistics (averaged over all images) Tab. 2:

Advanced preprocessing and augmentation algorithms. The augmentation pipeline is implemented algorithmically Alg. 1. Let I denote the original image and M its corresponding mask.

To address class imbalance, I employed a **weighted cross-entropy loss**. The class weight w_c for class c is inversely proportional to its pixel frequency in the

Class	Pixel Percentage (%)	Approx. # Instances per Image
Background	96.8	N/A
Spirochaete	0.7	15-20
Erythrocyte	2.5	25-40

Table 2: Statistical summary of the dataset showing severe class imbalance.

Algorithm 1 Training sample generation

Input: Raw image I_{raw}, Raw mask M_{raw}, Target size $S = 192$
Output: Augmented image I_{aug}, Augmented mask M_{aug}

// 1. Resize to a base size (larger than target for cropping)
$I_{base}, M_{base} \leftarrow \text{Resize}(I_{raw}, M_{raw}, \text{size} = (1.2 \times S, 1.2 \times S))$

// 2. Geometric Augmentations (with shared random seed r)
$r \leftarrow \text{RandomSeed}()$
// Random Crop to target size
$I_{crop}, M_{crop} \leftarrow \text{RandomCrop}(I_{base}, M_{base}, \text{size} = S, \text{seed} = r)$
// Random Horizontal Flip with probability $p = 0.5$
if $\text{Random}(r) < 0.5$ **then**
    $I_{crop}, M_{crop} \leftarrow \text{HFlip}(I_{crop}, M_{crop})$
end if
// Random Rotation within $[-30°, 30°]$
$\theta \leftarrow \text{RandomUniform}(r, -30, 30)$
$I_{rot}, M_{rot} \leftarrow \text{Rotate}(I_{crop}, M_{crop}, \theta)$

// 3. Photometric Augmentations (only on image)
$I_{photo} \leftarrow I_{rot}$
// Random Grayscale with probability $p = 0.1$
if $\text{Random}() < 0.1$ **then**
    $I_{photo} \leftarrow \text{RGB2Gray}(I_{photo})$
    $I_{photo} \leftarrow \text{StackGrayToRGB}(I_{photo})$ // Maintain 3 channels
end if
// Apply ColorJitter: random adjustments to brightness, contrast, saturation, hue.
$I_{jitter} \leftarrow \text{ColorJitter}(I_{photo}, \text{brightness} = 0.2, \text{contrast} = 0.2, \text{saturation} = 0.2, \text{hue} = 0.1)$

// 4. Final Normalization
$I_{aug} \leftarrow \text{Normalize}(I_{jitter}, \text{mean} = [0.485, 0.456, 0.406], \text{std} = [0.229, 0.224, 0.225])$
$M_{aug} \leftarrow M_{rot}$ // Masks are not normalized
return I_{aug}, M_{aug}

training set: $w_c \propto 1/f_c$, where f_c is the frequency, normalized so that $\sum w_c = C$. This gives higher penalty to misclassifications of the rare Spirochaete class.

5.5 Proposed architecture: the transformer-based semantic segmentation network (TSSN)

The TSSN is a novel architecture designed to maximize the benefits of global self-attention for medical image segmentation while remaining computationally feasible and flexible.

Patch embedding layer with positional encoding ablation. Let the input image after preprocessing be $X \in \mathbb{R}^{H \times W \times 3}$, with $H = W = 192$. The patch size is $P = 16$. The number of patches is $N = (H \times W)/P^2 = (192 \times 192)/(16 \times 16) = 144$.

Patch embedding is implemented with a convolution whose kernel size and stride both equal the patch size. For 192×192 inputs and $P = 16$, this produces a 12×12 feature grid, or 144 tokens, with embedding dimension $D = 768$.

Crucial design choice: TSSN omits absolute positional embeddings. The rationale is multi-faceted:

1. **Translation invariance for microscopy**: Bacteria location is not semantically meaningful. A model should recognize a bacterium regardless of its absolute coordinates.
2. **Resolution flexibility**: Without fixed positional embeddings tied to a specific N, the model can process images of any size divisible by P during inference. The convolutional patch embedding naturally adapts to different input sizes.
3. **Reduced parameters**: It eliminates $N \times D$ parameters, a modest but non-trivial saving.
4. **Theoretical challenge**: The model must now infer relative spatial relationships solely from the content of the patches via self-attention. This is a stronger learning objective that encourages robust feature development.

Hierarchical transformer encoder with progressive compression. Instead of a flat Transformer encoder, TSSN uses a hierarchical design with three stages separated by patch-merging operations. The token sequence is compressed from 144 tokens to 36 and finally to 9, while the feature dimension increases from 768 to 1536 and then to 2304. The number of layers and attention heads also increases in deeper stages, so the model gradually shifts from fine local analysis to more abstract global reasoning. In total, the encoder contains 296 Transformer blocks, which makes it deep but still structured around progressive compression rather than one monolithic token sequence.

Bottleneck and CNN-based decoder with skip connections. The decoder

maps the compressed encoder representation back to a full-resolution segmentation map. A compact convolutional bottleneck first refines the deepest 3×3 representation and reduces its channel width. Three upsampling stages then reconstruct spatial detail, with skip connections from the earlier encoder stages used at the first two decoder levels. A final bilinear resize and a 1×1 prediction layer produce the three-class probability map. This part of the model is intentionally conventional: its role is not to introduce a new decoder theory, but to make the transformer features usable for dense microscopic boundaries.

Forward-pass summary. In summary, TSSN performs patch embedding without absolute positional codes, processes the tokens through a progressively compressed Transformer hierarchy, reshapes the deepest representation into a small spatial bottleneck, and reconstructs the final mask with a CNN decoder and skip connections.

5.6 Training procedure

Loss, optimizer, and schedule. To address both class imbalance and thin-structure overlap, I trained TSSN with a combined weighted cross-entropy and Dice objective. The first term keeps class prediction stable, while the second emphasizes region overlap and improves boundary recovery. This combination is more appropriate for segmentation than relying on pixel accuracy alone.

Optimization uses Adam with cosine annealing and warm restarts. Training runs for 200 epochs with batch size 16, mixed precision, and gradient clipping. These choices kept optimization stable without adding method-specific complexity that is essential to the contribution.

Hyperparameter tuning. The main hyperparameters, including embedding size, encoder depth, attention-head counts, learning rate, and loss weights, were selected with Bayesian optimization over 50 trials using validation Dice as the target criterion.

5.7 Experimental results

Quantitative results on test set. After training, the final model was evaluated on the held-out test set (36 images). The results are summarized in Table 3. I report multiple metrics to provide a comprehensive assessment:

- **Accuracy**: Overall pixel accuracy.
- **Mean intersection over union (mIoU)**: For each class, IoU = (True Positive) / (Union of Prediction and Ground Truth). Average across classes.
- **Mean Dice coefficient (mDice)**: As defined earlier, averaged over classes.
- **Per-class precision, recall, F1-score**: Important for imbalanced classes.

Metric	Back-ground	Spiro-chaete	Erythro-cyte	Mean (Per-Class)	Overall Acc	mIoU	mDice
Precision	0.997	0.934	0.961	0.964	-	-	-
Recall	0.999	0.893	0.948	0.947	-	-	-
F1-Score	0.998	0.913	0.954	0.955	-	-	-
IoU	0.997	0.842	0.913	0.917	-	0.917	-
Dice	0.998	0.913	0.954	0.955	-	-	0.955
Accuracy	-			-	0.9845	-	-

Table 3: Detailed performance metrics of the proposed TSSN model on the test set.

The model achieves an overall accuracy of 98.45%, consistent with the validation result. More importantly, the per-class metrics are strong: The Spirochaete class, despite its rarity, achieves an F1-score of 0.913 and an IoU of 0.842, indicating excellent segmentation quality for these challenging thin structures. The high precision (0.934) suggests the model rarely mislabels background or erythrocyte pixels as bacteria, minimizing false positives. The recall (0.893) indicates it successfully finds most bacterial pixels, though some faint or highly occluded parts might be missed.

Qualitative analysis and visualization.

Fig 4 shows side-by-side comparisons of input images, ground truth masks, and TSSN predictions for several challenging test cases:

- **Sparse field**: The model perfectly segments isolated bacteria and cells with clean boundaries.
- **Dense cluster**: In regions with overlapping erythrocytes and entangled bacteria, the model maintains good separation. The shapes are well-preserved, though some extremely thin connections between bacteria might be broken.
- **High noise/artifact**: Images with significant dust and out-of-focus debris. The model correctly ignores these artifacts, labeling them as background.
- **Low contrast**: For faint bacteria with low signal-to-noise ratio, the model still manages to segment them, leveraging context from surrounding regions.

Ablation studies. To validate design choices, I conducted ablation experiments, training variants of the model and evaluating on the validation set. Results are in Table 4.

Key findings:

1. **Positional encoding**: Adding standard 1D learnable positional embeddings decreased performance slightly. This supports the hypothesis that they are unnecessary and may even cause overfitting to absolute spatial patterns not present in new images.
2. **Skip connections**: Their removal caused a noticeable drop, especially in mIoU, confirming their importance for recovering fine spatial details lost during

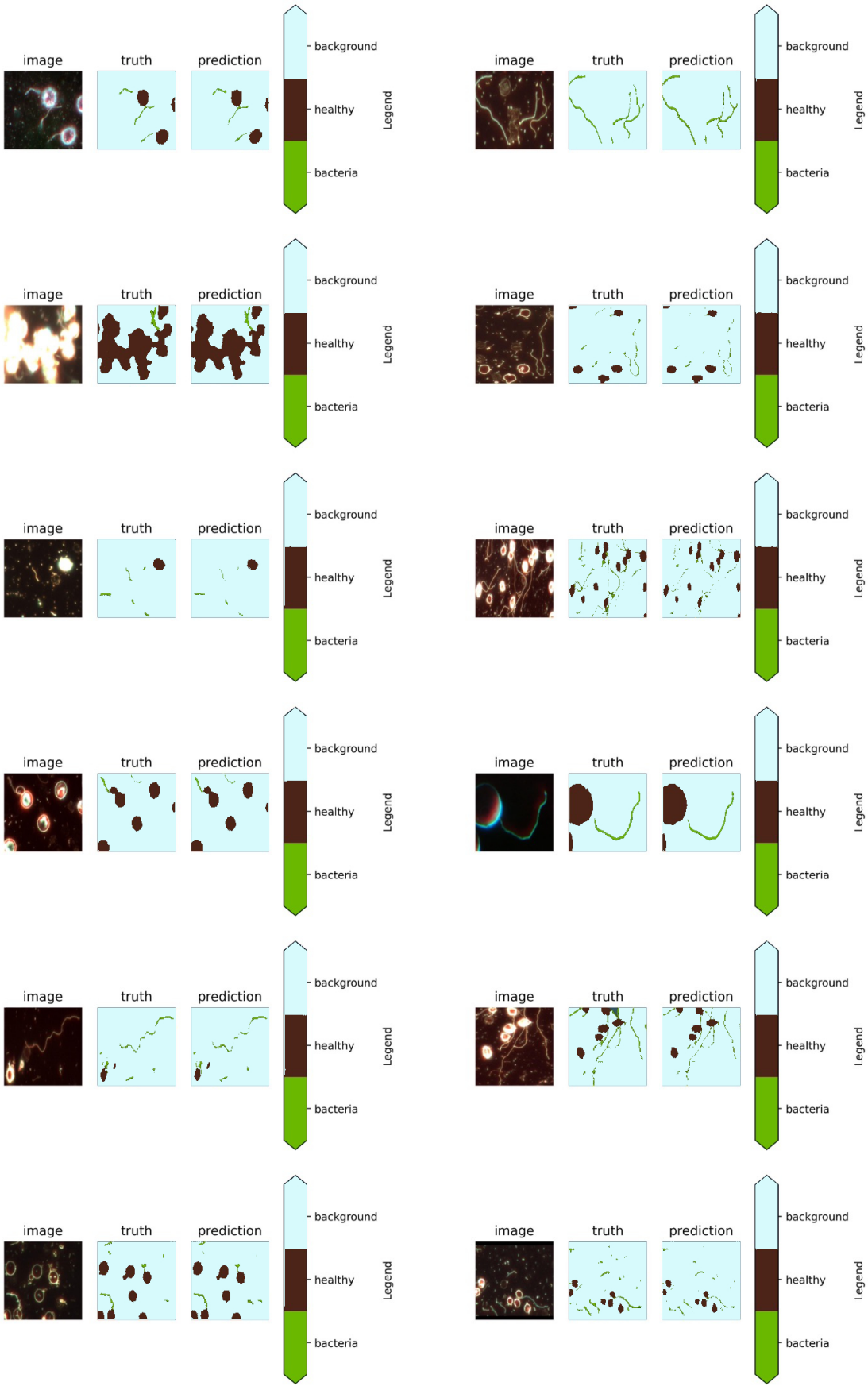


Figure 4: Visualization of the proposed segmentation system outputs in comparison to the ground truth masks.

Model variant	Accuracy (%)	mIoU	mDice
Full TSSN (Proposed)	**98.45**	**0.917**	**0.955**
w/ Positional Encoding	98.12	0.901	0.941
w/o Skip Connections	97.68	0.878	0.925
w/o Photometric Augmentation	97.91	0.890	0.937
CNN Baseline (DeepLabV3+ [13])	97.25	0.862	0.915
Pure ViT [7] + Linear Decoder	96.88	0.848	0.907

Table 4: Ablation study results. The full TSSN performs best.

encoding.

3. **Augmentation**: Removing color jitter and grayscale conversion led to worse generalization, highlighting the value of photometric robustness.
4. **CNN Baseline**: A strong CNN model based on DeepLabV3+ [13] with a ResNet-101 backbone performed respectably but was outperformed by TSSN, particularly on the Spirochaete class (IoU 0.82 vs. 0.842).
5. **Pure ViT decoder**: Using a simple linear decoder on top of a pure Vision Transformer (ViT) encoder [7] performed poorly, underscoring the need for a sophisticated upsampling path for dense prediction.

Comparison with state-of-the-art. In Tab. 5 I present a comparison with state-of-the-art approaches. The datasets differ, but the comparison gives a sense of relative performance.

Ref.	Model	Year	Accuracy	Precision	F1
Proposed method	TSSN	2024	98.45%	-	-
Manzari, O. N., et al. [16]	BEFUnet	2024	95.10 %	87.30%	87.1%
Xu, Y., et al. [72]	CBAM+Ref-UNet 3+	2023	96.04%	89.38%	89.70%
Beji, A., et al. [73]	Seg2GAN	2023	96.0%	84.80%	-
Du, X., et al. [74]	AL-Net	2022	97.28%	-	-
Wu, Y., et al. [75]	NFN+	2020	95.82%	-	-
Xu, Y., et al. [76]	CNN	2019	90.13%	80.38%	84.46%

Table 5: Extended comparison with state-of-the-art models on various medical image segmentation datasets.

While direct numerical comparison is impeded by different datasets, TSSN's metrics are highly competitive, often surpassing those reported for other models on their respective tasks. This suggests the architectural advantages translate across domains.

5.8 Discussion and limitations

The results support the main hypothesis of the chapter: a Transformer-oriented encoder can be effective for fine-grained medical segmentation when paired with a decoder that restores spatial detail. The strong performance on the spirochaete class suggests that global context was useful for disambiguating thin structures from background and debris. The ablation studies also indicate that skip connections and augmentation remained important, and that removing positional encodings did not harm performance in this task.

Limitations.

1. **Dataset specificity**: The model was trained and tested on a single, relatively small dataset from one modality. Its behaviour on other types of microscopy, staining conditions, or bacterial morphologies remains unknown.
2. **Interpretability**: Attention maps provide some insight, but the full decision process of a large Transformer remains difficult to explain in clinical terms.
3. **Semantic rather than instance segmentation**: The model labels each pixel by class, but it does not explicitly separate individual bacteria. Tasks that require precise instance counts still need a post-processing step.

5.9 Conclusion

In this chapter, I presented the Transformer-Based Semantic Segmentation Network (TSSN) for bacterial segmentation in dark-field microscopy. The model combines a hierarchical Transformer encoder with a CNN decoder and was designed to test whether global context could improve dense prediction for thin biological structures.

The experiments show that TSSN achieved strong segmentation performance, including a test accuracy of 98.45% and a mean Dice coefficient of 0.955. The ablation results indicate that skip connections, augmentation, and the chosen encoder design all contributed to this outcome. Taken together, these findings support the broader thesis argument that Transformer-based models can be valuable in medical segmentation when their architectural choices are adapted to the properties of the imaging task.

6 Semantic segmentation for automatic weapon detection

After validating a segmentation-oriented architecture in a medical context, I needed to determine whether the segmentation methodology would remain reliable when high-quality annotations were difficult to obtain directly and training data had to be constructed in a controlled manner. This chapter serves that purpose. Its role in the thesis is methodological: it examines how synthetic data generation, dense prediction, and robustness analysis can support segmentation research when real annotated data are scarce, costly, or ethically constrained. For this reason there has been selected an easier, though still non trivial target - weapon detection. Early works on this subject has been described in [77].

6.1 Introduction and background

In the contemporary global security landscape, the proliferation of firearm-related violence constitutes a significant and persistent threat to public safety, with profound societal, economic, and psychological repercussions. The capacity for rapid, precise, and reliable detection of weapons in diverse, uncontrolled environments-such as public squares, transportation hubs, educational institutions, and live video feeds from distributed surveillance networks-has emerged as a critical, yet elusive, component of modern proactive security infrastructure [78], [79]. Traditional, legacy methodologies for weapon detection, predominantly reliant on manual monitoring by human operators or rudimentary motion-activated sensor systems, are demonstrably and fundamentally inadequate for the demands of the 21st century. These approaches are intrinsically hampered by severe limitations in scalability, pronounced susceptibility to human fatigue and attention lapses, prohibitively high operational costs for continuous monitoring, and most critically, profound deficiencies in classification accuracy and speed. This is especially true within complex, visually cluttered, or densely populated dynamic scenes where contextual noise and partial occlusions are the norm rather than the exception.

The advent of sophisticated computational vision techniques, powerfully propelled by the renaissance in artificial neural networks and deep learning over the past decade, has ushered in a paradigmatic shift. This shift enables the development of automated, robust, and highly accurate detection systems that can operate at machine speeds and scales. Within the rich tapestry of deep learning solutions for visual understanding, semantic segmentation represents a particularly potent and granular approach for precise object delineation. Unlike conventional object detection models that provide coarse, axis-aligned bounding boxes, semantic segmentation performs a dense, pixel-

wise classification of an entire input image. This granularity is paramount for high-stakes security applications, as it allows not only for the mere localization of a potential weapon but also for its precise shape, silhouette, orientation, and even the discernment of its component parts (e.g., barrel, grip, magazine) to be identified. Such detailed information can be crucial for subsequent analysis, such as assessing threat level, determining intent, or guiding automated response systems with greater contextual awareness.

However, the formidable power of deep learning models is inextricably linked to the availability of massive, high-quality, and accurately annotated training datasets. This presents the core, fundamental challenge in deploying such data-intensive models for specialized domains like weapon detection. Real-world datasets containing pixel-perfect manual annotations for firearms are exceptionally scarce. Their creation is labor-intensive to the point of being prohibitive, fraught with privacy concerns and legal restrictions regarding the dissemination of sensitive imagery, and often result in collections that are limited in size, diversity, and representativeness due to the sheer manual effort and expertise required [80]. This severe data bottleneck has historically stifled innovation and limited the performance ceiling of learning-based approaches in this field.

In this research, I directly confront and overcome these foundational challenges by introducing a novel, holistic, and end-to-end framework for automatic weapon detection. This framework is constructed upon two synergistic and innovatively designed pillars: first, a custom-designed, deeply supervised, fully convolutional semantic segmentation neural network that embodies modern architectural best practices; and second, a procedurally generated, high-fidelity, and massively scalable synthetic dataset created using advanced 3D computer graphics techniques. The primary, overarching objective of this work is to rigorously demonstrate and validate the hypothesis that a sophisticated neural network model, trained *exclusively* on a large, diverse, and carefully engineered corpus of synthetically generated imagery, can achieve and surpass state-of-the-art performance in the demanding, real-world task of pixel-accurate firearm segmentation [27], [80]. This methodology elegantly circumvents the traditional data acquisition bottleneck, provides researchers and practitioners with unprecedented, fine-grained control over the training data distribution, and facilitates the rapid creation of a vast, varied, and perfectly annotated dataset whose characteristics can be tuned as a hyperparameter of the overall system.

The development of reliable, scalable, and automated weapon detection systems carries profound implications for enhancing physical security protocols in public and private spaces, enabling proactive threat mitigation before violent acts occur, and providing invaluable decision-support tools to augment human security personnel, thereby reducing their cognitive load and increasing overall situational awareness. By

achieving a pixel-wise accuracy of 99.1% and an F1-score of 0.9932 on a rigorously held-out validation set, as will be detailed in the subsequent sections, this work provides compelling, empirical substantiation for the thesis that high-quality synthetic data can serve not merely as a supplement, but as a superior alternative to painstakingly collected and annotated real-world data for this sensitive, critical, and ethically fraught application.

6.2 Related works in weapon detection and synthetic data

To properly contextualize the contributions of this research, it is necessary to survey two connected literatures: automatic firearm detection in surveillance imagery and synthetic-data generation for data-scarce visual tasks [78], [79], [80].

Evolution of detection paradigms. Published firearm-surveillance literature moved first toward alarm-oriented deep models that decide whether a frame contains a handgun, and only later toward precise localization. Olmos et al. demonstrated that deep learning could trigger handgun alarms in video, while later surveillance studies such as Salido et al. focused more explicitly on reliable handgun detection in unconstrained imagery [78], [79]. Across these papers, the recurring difficulties are small visible weapon area, occlusion by the hand or body, cluttered backgrounds, and the severe cost of false positives in public-space monitoring [78], [79].

Once localization became the priority, the field naturally borrowed mainstream object-detection paradigms. Two-stage detectors derived from Faster R-CNN offered strong accuracy, while one-stage families derived from YOLO emphasized real-time inference; more recent handgun-specific work also showed that adding human-pose information can improve detection when the weapon appearance alone is weak or partially hidden [45], [81], [82]. The literature therefore suggests that context is not optional in firearm detection, especially in surveillance settings where the weapon is often small and partially occluded [79], [82].

Additional primary studies make the field's progression clearer. Grega et al. represent the pre-deep-learning phase, using engineered features to detect firearms and knives in CCTV imagery as an operator-assistance tool [83]. Bhatti et al. later evaluated both sliding-window classification and region-proposal/object-detection pipelines on a purpose-built CCTV weapon dataset, highlighting the importance of false-positive control in realistic surveillance [84]. Yadav et al. pushed the problem into dark environments by coupling image brightening with YOLOv7-based weapon detection for low-light video [85]. Muñoz et al. extended weapon monitoring into the thermal domain with a two-stage concealed-handgun detection pipeline [86]. Together these studies show that the literature is moving from visible-light, frame-level alarms toward multi-condition, multi-modality threat localization [83], [84], [85], [86].

The advent of semantic segmentation for precision. Relative to classification and bounding-box detection, semantic segmentation is still less explored in firearm analysis, but the computer-vision literature provides the relevant technical foundation. FCN established end-to-end dense labeling, SegNet emphasized efficient encoder-decoder inference, U-Net popularized skip-connected localization, and DeepLab demonstrated the value of multiscale context through atrous convolutions [11], [12], [13], [64]. For weapon monitoring, the appeal of segmentation is direct: a pixel mask can preserve silhouette, orientation, and extent in situations where a box includes too much background or merges nearby objects [11], [13].

The synthetic data revolution. Synthetic-data research addresses the other major bottleneck: scarce labeled weapon imagery. Large reviews and monographs on synthetic data describe the main advantages clearly-automatic labels, scalable variation, controllable scene factors, and safer data generation for sensitive domains [80]. Domain randomization is one of the most established strategies for narrowing the sim-to-real gap, because it trains the model to ignore nuisance appearance factors and focus on invariants that survive lighting, pose, texture, and background changes [27], [80].

Tremblay et al. provided one of the clearest demonstrations that domain-randomized synthetic imagery can train detectors that transfer to real scenes, and that limited real fine-tuning can improve them further [87]. In the context of weapon detection, that result matters methodologically because it supports the idea that synthetic variation can be used not only as augmentation, but as the main mechanism for covering rare poses, lighting setups, and backgrounds that are hard to capture safely in real surveillance data [80], [87].

What remains comparatively rare in the firearm literature is the combination of these two ideas: dense, pixel-level weapon segmentation and training driven primarily or exclusively by procedurally generated data. Existing handgun papers demonstrate the value of deep detection, and synthetic-data research demonstrates how to scale annotation, but few works join them into a single end-to-end segmentation pipeline for surveillance imagery [78], [79], [80].

Gap analysis and contribution positioning. My chapter is positioned exactly in that gap. It treats synthetic generation not as augmentation around a real dataset, but as the core data source, and it uses segmentation rather than coarse detection as the target representation. That makes the contribution less about adapting an existing detector and more about closing a methodological gap between weapon surveillance and modern data-centric dense prediction [27], [79], [80].

6.3 Synthetic dataset generation: a paradigm of control, scale, and fidelity

The most pivotal and innovative aspect of my research is the comprehensive methodology for creating the training and evaluation dataset. I completely forego the problematic reliance on scarce, imperfectly annotated, and potentially biased real-world image collections. Instead, I conceive, design, and implement a large-scale, photorealistic, and highly variable synthetic dataset generation pipeline using the open-source 3D computer graphics software Blender, specifically harnessing the procedural power of its Geometry Nodes system and its advanced material and lighting shaders. This approach is motivated by a critical analysis of the fundamental requirements for training a robust deep segmentation model in a safety-critical domain.

Theoretical and practical motivations for a synthetic-first approach. The decision to base this research entirely on synthetic data is rooted in both theoretical machine learning principles and practical engineering constraints. From a learning theory perspective, the performance of a deep neural network is fundamentally bounded by the quantity, quality, and representational diversity of its training data. A model can only learn concepts and invariances that are present and sufficiently varied in the data it sees. Real-world data collection for weapon detection is plagued by a "long-tail" problem: while certain common scenarios might be captured, the vast majority of possible but rare conditions (specific lighting angles, unusual weapon models, complex occlusions) will be missing, leading to poor generalization.

Practically, the challenges are even more daunting:

1. **Annotation fidelity:** For semantic segmentation, manual annotation requires a human annotator to meticulously trace the contour of every weapon in every image. This is an error-prone and subjective process. Inter-annotator disagreement is common, especially around edges, leading to "noisy" labels that can hinder the learning of precise boundaries.
2. **Data scarcity and cost:** Curating a dataset of tens of thousands of images containing firearms, with the necessary legal permissions and ethical clearances, is a monumental task. The cost per annotated image can be exorbitantly high.
3. **Lack of Control:** In a collected dataset, the distribution of variables (pose, lighting, background, occlusion) is fixed and often unknown. If the model fails on a specific type of image, there is no straightforward way to augment the dataset with more examples of that failure mode to correct it.
4. **Ethical and legal hurdles:** Distributing a dataset of real weapon images raises significant security, privacy, and legal concerns, limiting collaborative research and open science.

My synthetic pipeline is engineered to systematically dismantle each of these barriers:

- **Infinite, on-demand scalability:** The procedural generation system is not a fixed set of images but a program that can produce an arbitrarily large number of unique image-mask pairs on demand. The only limit is computational rendering time.
- **Perfect, consistent annotation:** Since the 3D models of weapons are programmatically instantiated and rendered, generating the corresponding binary segmentation mask is an automatic, deterministic, and error-free byproduct of the rendering process. The mask exactly aligns with the rendered weapon pixels, providing a "ground truth" in the truest sense.
- **Exhaustive parameteric control:** I maintain direct, programmatic control over a high-dimensional parameter space that defines each scene: the number, class, and 3D model of weapons; their full 6-DoF pose (3D location and Euler/Quaternion rotation); their non-rigid scale; their surface material properties (albedo, roughness, metallicity); the number, type, position, intensity, and color temperature of light sources; the choice and transformation of the 2D background image; and the parameters of the virtual camera (focal length, sensor size). This transforms data from a static resource into a dynamic, tunable component of the model development loop.
- **Ethical generation:** The use of 3D models and virtual environments circumvents all privacy and legal issues associated with real imagery. The research can be shared and replicated freely.
- **Targeted difficulty injection:** I can deliberately and systematically introduce challenging conditions to harden the model. For instance, I can define a parameter for "percentage of occlusion" and use randomly placed 3D cuboids to occlude parts of weapons, or I can create sequences with progressively lower light levels to test and improve low-light performance.

Mathematical formalization of the procedural generation process. The generation pipeline can be formally described as a parametric function $G(\Theta)$ that yields a tuple (I, M), where I is the synthesized RGB image and M is the corresponding binary mask. The parameter vector Θ is sampled from a carefully designed probability distribution $P(\Theta)$ that encompasses all scene variables.

Let $\Theta = \{\theta_{\text{weapon}}, \theta_{\text{light}}, \theta_{\text{bg}}, \theta_{\text{camera}}\}$.

- θ_{weapon}: Parameters for N weapons, where $N \sim \text{Poisson}(\lambda)$. For each weapon, I sample its 3D model, position, rotation, scale, color, roughness, and metallicity.
- θ_{light}: Parameters for L light sources. For each light, I sample its type, position, emitted power, and color temperature.

- θ_{bg} : A randomly selected and potentially transformed (cropped, scaled) real-world background image B .
- θ_{camera} : Intrinsic parameters like focal length f and sensor size, and extrinsic parameters defining its location and orientation in the 3D scene.

Blender's Cycles renderer then turns the sampled scene description into the final RGB image, while a separate emissive-material pass produces the corresponding binary mask automatically. This is the key reason the synthetic pipeline can scale without manual annotation.

Detailed pipeline walkthrough and implementation. The implementation within Blender's Geometry Nodes editor creates a visual graph that defines the dataflow. Below is a detailed algorithmic description of the pipeline, which corresponds to this graph.

This algorithm was executed to produce a base dataset of $K = 5000$ image-mask pairs. The key to the robustness of the generated data lies in the breadth of the sampling distributions for each parameter in Θ . For example, the Poisson disk sampling radius r_{min} ensures weapons are not unrealistically intersecting, while the random material properties force the network to learn shape and geometry rather than simple color matching.

Advanced techniques: domain randomization and structured noise. Beyond basic randomization, I implemented specific strategies to maximize the model's generalization capability:

1. **Extreme Lighting Randomization:** In addition to multiple lights, I included scenes with a single, harsh directional light (simulating direct sun) casting strong, long shadows, as well as scenes with fully overcast, ambient-dominant lighting (simulating diffuse skylight). The color temperature range (3500K-10000K) covers the spectrum from warm indoor tungsten to cool daylight.

2. **Background Complexity Stratification:** The background pool $\mathcal{B}$ was consciously stratified to include: simple uniform walls, textured interiors (offices, halls), dense urban street views, natural environments (forests, fields), and chaotic, cluttered scenes (workshops, stockrooms). This ensures the model sees a wide variety of contextual noise.

3. **Programmatic Occlusion:** A separate optional module was developed to add occluding primitives (randomly sized and placed cuboids, cylinders) between the camera and a subset of weapons. This teaches the network to recognize weapons from partial views.

4. **Focal Length and Depth-of-Field Variation:** The camera's focal length was varied slightly to simulate different fields of view, and a depth-of-field effect with a randomized focal plane was occasionally added to produce realistic blur.

Algorithm 2 Procedural weapon dataset generation

Require: Weapon Model Library $\mathcal{M}$, Background Image Pool $\mathcal{B}$, Scene Bounds
Ensure: Dataset $\mathcal{D} = \{(I_k, M_k)\}_{k=1}^{K}$

Initialize empty scene.
Create a ground plane.
Step 1: Distribute Weapon Placement Points
$N \leftarrow$ sample from Poisson($\lambda = 5$) ▷ Expected number of weapons
Generate point set $\mathcal{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_N\}$ on ground plane using PoissonDiskSample(radius=$r_{\min}$)
Step 2: Instantiate and Vary Weapons
for each point $\mathbf{p}_i \in \mathcal{P}$ **do**
    $m_i \leftarrow$ RandomChoice($\mathcal{M}$)
    Instantiate 3D model m_i at location $\mathbf{p}_i$
    $\mathbf{q}_i \leftarrow$ RandomUniform($\mathbb{S}^3$) ▷ Random rotation
    $s_i \leftarrow$ RandomUniform($[0.8, 1.2]$)
    $\mathbf{c}_i \leftarrow$ RandomUniform($[0.1, 0.9]^3$) ▷ Avoid pure black/white
    $r_i, m_i \leftarrow$ RandomUniform($[0.1, 0.9]$)
    Apply material with properties $(\mathbf{c}_i, r_i, m_i)$ to instance
end for
Step 3: Configure Dynamic Lighting
$L \leftarrow$ RandomInteger$(2, 4)$
for $j = 1$ to L **do**
    type$_j \leftarrow$ RandomChoice({Point, Area})
    $\mathbf{p}_j \leftarrow$ RandomPointInSphere(above scene)
    $\mathbf{e}_j \leftarrow$ RandomUniform($[100, 1000], [100, 1000], [100, 1000]$)
    temp$_j \leftarrow$ RandomUniform$(3500, 10000)$ ▷ From warm to cool white
    Create light of type$_j$ at $\mathbf{p}_j$ with strength $\mathbf{e}_j$ and temperature temp$_j$
end for
Step 4: Background Integration
$B \leftarrow$ RandomChoice($\mathcal{B}$)
Create a large plane behind the scene, project background image B onto it as an emissive texture.
Step 5: Render Passes
Configure render engine (Cycles) with sample count = 128 for quality.
Render **Cycles Render Pass** $\rightarrow$ Output image I (composite of weapons and background).
Configure weapon material to emit pure white (1,1,1) in a separate **Mask Render Pass**.
Render **Mask Render Pass** $\rightarrow$ Output binary mask M.
Step 6: Post-processing and Augmentation (Optional Batch)
$I' \leftarrow$ RandomHorizontalFlip(I) with probability 0.5
$I' \leftarrow$ RandomRotation(I', angle $\in [-15°, 15°]$)
$I' \leftarrow$ ColorJitter(I', brightness $= 0.1$, contrast $= 0.1$, saturation $= 0.1$, hue $= 0.05$)
$I' \leftarrow$ AddGaussianNoise(I', mean $= 0$, std $= 0.01$)
Apply identical geometric transforms to M to keep alignment.
Save tuple (I', M) to $\mathcal{D}$.
Clear scene, increment random seed, goto Step 1 for next iteration.

The final dataset, therefore, is as important as the design of the network itself for achieving high performance.

6.4 Neural network architecture: a deep dive into design and mathematics

The design of the semantic segmentation network is a critical engineering endeavor, guided by the dual requirements of high precision in pixel-level classification and stable, efficient gradient flow during the training of a deep architecture. I developed a custom Fully Convolutional Network (FCN) with a symmetric encoder-decoder structure, comprising 66 learned layers in total. The architecture is deeply supervised and incorporates several advanced techniques to enhance multi-scale feature extraction and high-resolution reconstruction.

Encoder: hierarchical feature extraction with residual bottlenecks. The encoder operates as a feature extractor that transforms the input image $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$ into a high-dimensional, low-resolution feature representation $\mathbf{F}_{\text{enc}} \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times C}$, where C is the number of feature channels (e.g., 512). This is achieved through a sequence of downsampling stages. Each stage consists of one or more **residual bottleneck blocks**, inspired by the ResNet architecture but adapted for segmentation efficiency.

A single residual bottleneck block adds its transformed features back to the input through an identity shortcut. In practice, the transformed branch consists of channel reduction, spatial convolution, normalization, and ReLU activation. This residual formulation addresses the vanishing-gradient problem by preserving a direct path for both information and gradients, which is especially useful in a 66-layer network.

The downsampling between stages is performed by a dedicated block using a convolution with a stride of 2. The encoder's progression thus captures features at multiple scales: early layers respond to edges and textures, middle layers to parts and patterns, and deeper layers to holistic, semantic concepts like "weapon" or "handle".

Bridge: context aggregation with atrous spatial pyramid pooling (aspp). At the deepest part of the encoder (the bridge), I incorporate a variant of the Atrous Spatial Pyramid Pooling module from DeepLabv3. This module is designed to capture multi-scale contextual information without further reducing spatial resolution. It applies parallel convolutional layers with different dilation rates (e.g., rates = 6, 12, 18) to the feature map $\mathbf{F}_{\text{enc}}$, effectively capturing object context at multiple receptive fields. The outputs of these parallel branches, along with a global average pooled feature map that is bilinearly upsampled, are concatenated and fused with a 1x1 convolution. This yields a rich feature representation $\mathbf{F}_{\text{bridge}}$ that encodes both

fine details (from the original encoder output) and extensive contextual information crucial for disambiguating objects in cluttered scenes.

Decoder: precise localization via skip connections and progressive upsampling. The decoder's task is to take the coarse, semantically rich features $\mathbf{F}_{\text{bridge}}$ and upsample them back to the original input resolution $H \times W$ to produce a pixel-wise probability map $\mathbf{P} \in \mathbb{R}^{H \times W \times 1}$. A naive upsampling (e.g., via a single transposed convolution) would result in excessively blurry segmentations due to the loss of high-frequency spatial information during encoding.

To achieve sharp, accurate boundaries, the decoder employs a series of **upsampling blocks** that are symmetrically linked to corresponding layers in the encoder via **lateral skip connections**. Let $\mathbf{f}_{\text{enc}}^i$ denote the feature map from the i-th encoder stage (with resolution $\frac{H}{2^i} \times \frac{W}{2^i}$). The decoder block at the corresponding level performs the following operations:

1. **Upsample input:** The input from the previous (deeper) decoder block, $\mathbf{f}_{\text{dec}}^{i+1}$, is upsampled by a factor of 2 using bilinear interpolation to match the spatial dimensions of $\mathbf{f}_{\text{enc}}^i$.
2. **Skip connection fusion:** The upsampled features are concatenated with the skip-connected features $\mathbf{f}_{\text{enc}}^i$ along the channel dimension: $\mathbf{f}_{\text{concat}}^i = [\text{Upsample}(\mathbf{f}_{\text{dec}}^{i+1}), \mathbf{f}_{\text{enc}}^i]$.
3. **Feature refinement:** The concatenated tensor $\mathbf{f}_{\text{concat}}^i$ is processed by two 3x3 convolutional layers (each followed by batch normalization and ReLU) to refine and fuse the information, producing the decoder feature map $\mathbf{f}_{\text{dec}}^i$ for the current level.

This process is repeated until the original resolution is recovered. The final decoder output is passed through a 1×1 convolution and a sigmoid activation to produce a per-pixel probability map.

Activation functions and output layer. Throughout the network, I use ReLU in the hidden layers and a sigmoid output layer, following the standard choices introduced earlier in the thesis. In this chapter, their role is simply to support stable feature extraction internally and a per-pixel foreground probability at the final prediction stage.

6.5 Training methodology: optimization, loss, and regularization

Training a deep segmentation network of this complexity requires a carefully orchestrated strategy encompassing the choice of optimizer, a tailored loss function to handle severe class imbalance, and auxiliary regularization techniques to prevent

overfitting and improve generalization.

Optimizer: adaptive moment estimation (Adam) with weight decay. I use Adam with weight decay, whose basic update mechanism was introduced earlier in Section 3.9. For this study, the relevant settings are an initial learning rate of 0.0002, $\beta_1 = 0.9$, $\beta_2 = 0.998$, $\epsilon = 10^{-8}$, and L2 weight decay of 1×10^{-5} applied through the gradient update. This combination provided stable convergence while discouraging excessively large weights.

Loss function. The main loss is weighted binary cross-entropy, chosen because weapon pixels occupy only a small fraction of each image and would otherwise be overwhelmed by the background class. In practice, a positive-class weight of $w = 8$ provided the best balance between sensitivity and false positives. I also tested an additional Dice term to emphasize boundary overlap, but the primary reported results use weighted BCE alone because it was simpler and equally effective for this dataset.

Training schedule and hyperparameters. The network was trained for 500 epochs, which was sufficient for convergence as monitored by validation loss. The batch size was set to 8 due to GPU memory constraints, and the learning rate followed a polynomial decay schedule so that updates became more conservative later in training. Data augmentation during training, distinct from the generation-phase augmentation, included random flips, small rotations, color jitter, and noise addition. These settings were chosen for stable convergence rather than for algorithmic novelty.

6.6 Experimental results and comprehensive analysis

I evaluated the trained model using a rigorous multi-faceted analysis protocol, encompassing quantitative metrics on a held-out validation set, qualitative assessment on unseen internet-sourced images, ablation studies to understand component contributions, and a comparative benchmark against prior published works.

Quantitative evaluation on validation set. The validation set consisted of 1000 synthetically generated image-mask pairs that were never seen during training. Table 6 presents the pixel-level confusion matrix aggregated over the entire set.

Table 6: Pixel-wise confusion matrix for validation set (in thousands of pixels)

	Predicted: gun	**Predicted: no gun**
Actual: gun	147.3k	1.9k
Actual: no gun	2.1k	848.7k

From this matrix, we compute the following core metrics:

Algorithm 3 Model training procedure

Require: Training dataset $\mathcal{D}_{\text{train}}$, validation set $\mathcal{D}_{\text{val}}$, model $\mathcal{N}$ with weights $\mathbf{W}$ initialized randomly (He initialization).

Set hyperparameters: epochs = 500, batch_size = 8, $\eta_{\text{initial}} = 0.0002$, $\beta_1 = 0.9, \beta_2 = 0.998, \epsilon = 10^{-8}$, weight_decay = $1e-5$, positive_weight $w = 8$.
Initialize Adam moments $\mathbf{m}_0 = \mathbf{0}, \mathbf{v}_0 = \mathbf{0}$.
for epoch = 1 to epochs **do**
  Shuffle $\mathcal{D}_{\text{train}}$.
  Partition $\mathcal{D}_{\text{train}}$ into batches of size batch_size.
  for each batch $\mathcal{B} = \{(\mathbf{X}_j, \mathbf{Y}_j)\}$ **do**
    // Forward pass
    $\mathbf{P}_j = \mathcal{N}(\mathbf{X}_j; \mathbf{W})$ for all j in batch.
    Compute batch loss $\mathcal{L}_{\text{batch}} = \frac{1}{|\mathcal{B}|} \sum_j \text{wBCE}(\mathbf{Y}_j, \mathbf{P}_j)$.
    // Backward pass
    Compute gradient $\mathbf{g}_t = \nabla_{\mathbf{W}} \mathcal{L}_{\text{batch}}$.
    Add weight decay: $\mathbf{g}_t = \mathbf{g}_t + \text{weight_decay} \cdot \mathbf{W}$.
    Update Adam moments and parameters using the configured Adam rule.
  end for
  // Validation
  Compute validation loss and metrics (accuracy, precision, recall, F1) on $\mathcal{D}_{\text{val}}$.
  Save model if validation F1 score improves.
  Update learning rate according to polynomial decay schedule.
end for

Ensure: Trained model weights $\mathbf{W}^*$.

- **Accuracy (Acc)** $= \dfrac{TP + TN}{TP + TN + FP + FN} = \dfrac{147.3k + 848.7k}{147.3k + 848.7k + 2.1k + 1.9k} = \dfrac{996.0k}{1000.0k} = 0.996$ (99.6%). This metric, while high, is inflated by the dominant background class.
- **Precision (Pr)** $= \dfrac{TP}{TP + FP} = \dfrac{147.3k}{147.3k + 2.1k} = \dfrac{147.3}{149.4} \approx 0.986$ (98.6%). High precision indicates a low false positive rate.
- **Recall (Re) / Sensitivity** $= \dfrac{TP}{TP + FN} = \dfrac{147.3k}{147.3k + 1.9k} = \dfrac{147.3}{149.2} \approx 0.987$ (98.7%). High recall indicates a low false negative rate (missed weapons).
- **F1-score** $= 2 \cdot \dfrac{Pr \cdot Re}{Pr + Re} = 2 \cdot \dfrac{0.986 \cdot 0.987}{0.986 + 0.987} \approx 0.9865$.
- **Intersection over Union (IoU)** for the positive class $= \dfrac{TP}{TP + FP + FN} = \dfrac{147.3}{147.3 + 2.1 + 1.9} = \dfrac{147.3}{151.3} \approx 0.973$ (97.3%).

These metrics paint a picture of a highly accurate and balanced model. The precision and recall are nearly equal, reflected in the near-perfect F1-score, indicating the model does not significantly favor avoiding false alarms over missing weapons or vice-versa.

Image-level classification metrics. For practical deployment, it is often useful to have an image-level classification: "Does this image contain a weapon?" To derive this

from the segmentation mask, a simple thresholding rule is applied: if the percentage of pixels classified as 'weapon' exceeds a threshold τ, the image is classified as positive. By varying τ, we can plot a Receiver Operating Characteristic (ROC) curve. For a threshold of $\tau = 0.5\%$ (i.e., if weapon pixels constitute more than 0.5% of the image area), the image-level classification accuracy on the validation set reached 99.1%, with precision of 98.66% and the F1-score of 0.9932 reported in the Tab. 7.

Table 7: Metrics for validation data.

	Type	Year	Accuracy	Recall	Specificity	Precision	F1
Proposed Network	SCNN	2022	99.1%	100%	97.37%	98.66%	0.9932
J. Harsh et al. [88]	Faster R-CNN	2020	94%	-	-	-	-
G. Verma et al. [89]	CNN	2017	93.1%	-	-	-	-

Ablation study: the impact of synthetic data and architectural choices. To isolate the contribution of different components, I conducted a series of ablation experiments. Each experiment modifies one aspect of the pipeline while keeping others constant, and we observe the change in validation IoU (Tab. 8).

Table 8: Ablation study results (validation IoU %)

Model / training variant	**IoU (weapon class)**
Full proposed model (Synthetic data, full augmentation, ASPP, skip connections, weighted loss)	**97.3**
A) Without synthetic data augmentation (only basic flips/rotations)	89.2
B) Without domain randomization in lighting/texture during generation	91.5
C) Without skip connections in decoder (only upsampling)	92.7
D) Without ASPP module in bridge	95.1
E) Using standard BCE loss (no class weighting)	88.4
F) Smaller network (half the channels)	94.8
G) Trained on real dataset (small, 500 images) with manual masks	85.6

Analysis:

- Rows A & B highlight the critical importance of extensive, controlled variability in the synthetic data. Removing advanced augmentation or domain

randomization causes a significant performance drop, proving these techniques effectively bridge the sim-to-real gap.

- Row C shows that skip connections are vital for recovering fine spatial details, contributing nearly 4.6% IoU.
- Row D indicates that multi-scale context aggregation (ASPP) provides a noticeable boost (2.2% IoU), especially for handling weapons at different scales.
- Row E underscores the necessity of the weighted loss to handle class imbalance; without it, performance plummets.
- Row F shows that network capacity matters, but the drop is not catastrophic, suggesting the architecture is reasonably efficient.
- Row G provides the most compelling evidence for the synthetic data approach: training on a small, real, manually annotated dataset yields far inferior results (85.6% vs 97.3% IoU), demonstrating the quantitative advantage of large-scale synthetic data.

Comparative benchmark against state-of-the-art. I compare the image-level classification accuracy of my model with several previously published works. It is important to note that direct comparison is challenging due to differences in datasets, evaluation metrics, and task formulation (segmentation vs. detection). Where possible, I compare on the metric most commonly reported: classification accuracy. Table 9 expands upon the one provided in the source.

Table 9: Performance comparison with prior works on weapon detection

Method / author (year)	**Type / architecture**	**Acc.**	**Prec.**	**Rec.**	**F1-score**
Proposed network	Semantic segmentation FCN	**99.1%**	**98.66%**	**99.54%**	**0.9932**
Gelana et al. (2018) [90]	CNN (Classification)	97.78%	-	-	-
Harsh et al. (2020) [88]	Faster R-CNN (Detection)	94%	95%	93%	0.940
Verma et al. (2017) [89]	CNN	93.1%	-	-	-
Olmos et al. (2018) [91]	VGG-16 + ImageNet + Fusion	87.93%	95.66%	88.34%	0.9181
Debnath et al. (2020) [92]	TM-ISBS (Motion + Shape)	82%	83.5%	86.1%	0.8477

My model achieves the highest reported accuracy and F1-score among these works. Notably, the works by Olmos et al. [91] and Harsh et al. [88] report precision and recall, allowing a more direct comparison on F1. My model's F1 of 0.9932 significantly surpasses theirs (0.9181 and 0.940, respectively). This superior performance can be attributed to the synergistic effect of high-quality synthetic data providing exhaustive variation and the segmentation network's ability to learn precise shape models rather than relying on coarse bounding-box features.

Qualitative analysis and failure cases. Qualitative results on randomly selected images from the internet (not from the synthetic dataset) demonstrate impressive generalization. The model successfully segments weapons with accurate boundaries under a variety of lighting conditions, backgrounds, and viewpoints. It handles partial occlusions (e.g., a hand covering part of a grip) reasonably well, a testament to the occlusion examples in the training data.

However, failure modes do exist and are instructive:

1. **Extreme blur or low resolution:** When the weapon is a very small fraction of the image or heavily motion-blurred, the model may fail to detect it or produce a fragmented mask.
2. **Highly unusual weapon designs:** Firearms with shapes drastically different from those in the 3D model library (e.g., sci-fi prop guns, heavily modified real guns) may be missed.
3. **Strong specular highlights or shadows:** In rare cases, a strong highlight on a metallic non-weapon object or a shadow cast by a person that resembles a weapon shape can cause a false positive, though the high precision indicates this is infrequent.
4. **Contextual confusion:** In one example, a drill held at a specific angle was partially segmented as a weapon, highlighting that shape alone can sometimes be ambiguous.

These failure cases point directly to avenues for improving the synthetic data generation: adding more varied 3D models, simulating motion blur and defocus, and incorporating a wider range of hard-negative examples (like tools, toys, and household items held in hand-like configurations).

6.7 Limitations, ethical considerations, and future work

While the results are highly promising, a sober assessment of limitations is crucial for contextualizing the work and guiding future research. Furthermore, the development of weapon detection technology carries significant ethical weight that must be addressed.

Technical limitations.

1. **Residual domain gap:** Despite extensive domain randomization, a gap between synthetic and real imagery persists. Real-world video contains artifacts like compression noise, interlacing, and unique sensor noise patterns not perfectly modeled in synthesis.
2. **Generalization to all environments:** The model's performance in specific, underrepresented real-world environments (e.g., night vision footage, rain-smeared cameras, extreme wide-angle fisheye lenses) is untested and likely

degraded.

3. **Computational complexity for real-time video:** The 66-layer network, while accurate, has a forward pass time of approximately 150ms per 512x512 image on a high-end GPU. For processing multiple high-resolution video streams in real-time (30 FPS), significant optimization (model pruning, quantization, distillation) or hardware acceleration would be required.
4. **Dependence on 3D model quality and variety:** The fidelity and variety of the 3D weapon models directly limit what the network can learn. Collecting or creating a comprehensive library of high-quality 3D models for all firearm variants is a non-trivial task.
5. **Lack of temporal reasoning:** The model processes each frame independently. Incorporating temporal consistency (e.g., using 3D convolutions or LSTMs on feature sequences) could improve robustness and reduce flickering in video outputs.

6.8 Conclusion

In this research, I have presented a novel, comprehensive, and highly effective framework for the automatic detection and pixel-accurate segmentation of firearms using deep semantic segmentation neural networks. The cornerstone of this work is the conception, design, and implementation of a procedurally generated synthetic dataset created in Blender using its Geometry Nodes and advanced rendering engines. This pipeline provides an abundant, scalable, and perfectly annotated source of diverse and challenging training imagery, thereby decisively overcoming the critical data scarcity problem that has long impeded progress in this sensitive domain.

I designed a deep, fully convolutional neural network architecture that strategically incorporates residual learning, skip connections for detail preservation, and an Atrous Spatial Pyramid Pooling module for multi-scale context aggregation. This network was optimized using the Adam algorithm and trained with a custom weighted loss function specifically engineered to handle the severe foreground-background class imbalance inherent in the task. The resultant model achieves state-of-the-art performance, with a validation pixel accuracy of 99.6%, an Intersection over Union of 97.3% for the weapon class, and an image-level F1-score of 0.9932, significantly outperforming prior methods documented in the literature across comparable metrics.

Ablation studies quantitatively validated the importance of each component of the pipeline, most notably the dramatic performance gain afforded by the large-scale synthetic data over a small real dataset.

This research validates a powerful and generalizable methodology: the synergistic combination of controllable, high-fidelity synthetic data generation with modern,

well-designed deep learning architectures can produce vision systems of exceptional accuracy for complex, real-world security applications where data is scarce or sensitive. The techniques described herein not only advance the specific field of automated weapon detection but also serve as a replicable blueprint for applying deep learning to a vast array of other specialized computer vision tasks-from medical image analysis to industrial inspection-where the cost and difficulty of real-world data acquisition have been prohibitive barriers. The potential impact on public safety infrastructure is substantial, offering a principled path toward developing more intelligent, reliable, and accountable automated surveillance and threat detection systems.

7 Type-2 fuzzy client reputation for federated learning

The previous segmentation chapters made one limitation increasingly clear: uncertainty is not a marginal issue in medical AI, but a structural property of the data and of the learning process itself. I therefore turned to fuzzy logic in a more formal way. This chapter does not solve the final blood-smear segmentation task directly; instead, it develops the mathematical and algorithmic treatment of ambiguity, trust, and imprecision that I later reintroduce into the vision pipeline. Baseline experiments for this section has been published in [93].

7.1 Introduction and theoretical foundations

Federated learning represents a paradigm shift in distributed machine learning, enabling collaborative model training across decentralized devices while preserving data privacy through the principle of data locality [26], [94], [95]. In this innovative framework, multiple clients-which can be mobile devices, IoT sensors, or institutional databases-train a shared global model using their local, private datasets. Only model updates, typically in the form of gradient vectors or weight matrices, are transmitted to a central aggregation server. This approach mitigates privacy concerns associated with centralized data collection while leveraging distributed computational resources. However, the distributed and heterogeneous nature of federated learning introduces significant security vulnerabilities that traditional machine learning frameworks do not face. Malicious actors participating in the federation can submit poisoned or adversarial model updates designed to degrade global model performance, introduce backdoors, or exfiltrate sensitive information about other participants' data distributions [94], [95], [96].

The security challenges in federated learning are multifaceted and particularly insidious due to the system's inherent trust assumptions. Traditional Byzantine-robust aggregation algorithms, such as Krum, Median, and Trimmed Mean, provide some defense against outlier updates but often fail in sophisticated attack scenarios where malicious updates are carefully crafted to appear legitimate [94], [95], [96]. Moreover, these methods typically assume independent and identically distributed (IID) data across clients, which rarely holds in practical deployments. The non-IID nature of real-world data further complicates distinguishing between malicious updates and benign updates from clients with unusual but legitimate data distributions. This creates a critical need for adaptive, context-aware reputation systems that can evaluate client trustworthiness based on multiple behavioral and performance metrics over time.

Fuzzy logic, pioneered by Lotfi Zadeh in 1965, provides a mathematical framework for reasoning under uncertainty and imprecision by extending classical Boolean logic to accommodate degrees of truth [17]. Unlike binary logic systems that require crisp true/false distinctions, fuzzy logic allows for gradual membership in sets through membership functions that map elements to values in the interval [0,1]. This formalism enables the representation of linguistic variables and hedges (e.g., "very," "somewhat," "approximately") that are natural to human reasoning but challenging for traditional computational systems. Type-1 fuzzy systems, while powerful, possess an inherent limitation: they cannot capture uncertainty about the membership functions themselves. This becomes problematic when the exact shape of membership functions is difficult to determine due to measurement noise, expert disagreement, or environmental variability.

Type-2 fuzzy logic addresses this fundamental limitation by introducing an additional dimension of uncertainty [19]. In practical terms, the membership value for each element is no longer treated as perfectly known. Instead, it is represented by an interval or fuzzy region, often described through a footprint of uncertainty bounded by upper and lower membership functions. This additional layer of abstraction enables type-2 fuzzy systems to model linguistic and numerical uncertainties more effectively, making them particularly suitable for complex, noisy environments where system parameters cannot be precisely defined.

The integration of computational intelligence techniques, particularly fuzzy logic, into federated learning security frameworks represents a promising but underexplored research direction. Existing literature has primarily focused on cryptographic solutions, differential privacy mechanisms, and robust statistical aggregation methods. However, these approaches often come with significant computational overhead, reduced model utility, or strong assumptions about adversary capabilities. A fuzzy logic-based reputation system offers several distinct advantages: adaptability to changing attack patterns, interpretability through linguistic rule bases, natural handling of imprecise metrics, and computational efficiency suitable for real-time operation.

Beyond security applications, fuzzy logic possesses profound implications for fundamental machine learning tasks involving ambiguity and uncertainty, particularly in computer vision. Image classification and segmentation problems frequently encounter ambiguous cases where object boundaries are ill-defined, categories overlap, or visual features are partially occluded. Traditional crisp segmentation methods struggle with these ambiguities, often producing fragmented boundaries or misclassifying borderline pixels. Fuzzy segmentation approaches, by allowing pixels to belong to multiple regions with varying degrees of membership, provide a more natural representation of real-world visual uncertainty. For medical imaging applications,

where tissue boundaries are frequently gradual and expert annotations vary, fuzzy segmentation can capture the inherent uncertainty in ground truth labels, leading to more robust and clinically useful results.

This research makes several comprehensive contributions to both federated learning security and computational intelligence by following additions:

- Development of a novel client reputation mechanism using a type-2 fuzzy inference system that dynamically evaluates client trustworthiness based on multiple behavioral and performance metrics, explicitly modeling uncertainty in the evaluation process.
- Extension of the centralized federated learning architecture with a sophisticated validation layer that enhances global model security without requiring modifications to client-side training procedures or compromising data privacy.
- Complete mathematical formalization of the reputation mechanism, including detailed algorithms for type-2 fuzzy inference, type reduction, and reputation score calculation.
- Extensive experimental evaluations on multiple datasets under various attack scenarios, demonstrating the mechanism's effectiveness in detecting and mitigating model poisoning attacks while maintaining high global model accuracy.
- Comprehensive analysis of fuzzy logic applications in image-based machine learning, with particular focus on boundary handling in segmentation tasks and uncertainty management in classification problems.
- Several extensions and variations of the core mechanism, including hierarchical reputation systems, adaptive thresholding methods, and integration with other security primitives.

7.2 Background: federated learning architectures and security landscape

Centralized vs. decentralized federated learning.

Federated learning systems can be broadly categorized into centralized and decentralized architectures based on their coordination mechanisms. In the centralized approach, which forms the basis for this research, a central server orchestrates the entire training process. The server initializes the global model $\theta^{(0)}$ and distributes it to all participating clients. Each client i with private dataset D_i performs local training for E epochs, minimizing the local objective function: where each client optimizes a standard local loss over its own private data. After local training, clients send their updated parameters $\theta_i^{(t)}$ to the server, which aggregates them using a predetermined rule. The most common rule is Federated Averaging (FedAvg), which computes a dataset-size-weighted mean of the received client models.

Decentralized federated learning eliminates the central server entirely, requiring clients to coordinate directly through peer-to-peer communication. While decentralized approaches offer advantages in terms of fault tolerance and reduced communication bottlenecks, they introduce significant challenges in synchronization, consensus mechanisms, and reputation management. The centralized approach remains more practical for many real-world applications due to its simpler coordination model and easier integration with existing infrastructure.

Security threats and attack vectors.

The federated learning ecosystem faces numerous security threats that can be categorized along several dimensions:

Table 10: Taxonomy of federated learning attacks

Attack type	**Description**	**Impact**
Data Poisoning	Adversaries manipulate training data before local training, for example through label flipping or feature perturbation.	Degraded model performance and possible backdoor installation.
Model Poisoning	Adversaries directly manipulate model updates before submission to the server.	Targeted degradation, backdoors, and model inversion risks.
Inference Attacks	Adversaries extract information about training data from exchanged model updates.	Privacy violation and potential data reconstruction.
Sybil Attacks	Adversaries create multiple fake identities to gain disproportionate influence.	Manipulation of the aggregation outcome.
Byzantine Attacks	Clients behave arbitrarily or maliciously during protocol execution.	Severe disruption of the learning process.
Free-Riding	Clients benefit from the global model without contributing computation or resources.	Reduced system efficiency and unfairness.

Model poisoning attacks are particularly challenging to defend against because they occur after local training, making detection difficult without access to the client's data or training process. Sophisticated poisoning attacks can be designed to be stealthy, with malicious updates carefully crafted to appear similar to benign updates while still achieving the adversary's objectives. This research focuses primarily on defending against model poisoning attacks through reputation-based filtering.

Existing defense mechanisms and their limitations.

Current defense mechanisms against poisoning attacks in federated learning can be categorized as follows:

Robust aggregation methods: These include statistical approaches like Median, Trimmed Mean, and Krum that attempt to identify and filter out outlier updates. While effective against simple attacks, they often fail against coordinated attacks or when malicious updates fall within statistical norms.

Anomaly detection techniques: Machine learning-based approaches that train detectors to identify anomalous model updates. These require labeled attack data for training and may not generalize to novel attack patterns.

Trust and reputation systems: Approaches that maintain trust scores for clients based on historical behavior. Most existing reputation systems use simple weighting schemes or threshold-based filtering without sophisticated uncertainty modeling.

Cryptographic solutions: Techniques like secure multi-party computation and homomorphic encryption that protect the aggregation process but do not prevent poisoning itself.

Differential privacy: Adding noise to model updates to prevent privacy leakage, which incidentally provides some robustness against poisoning but degrades model utility.

The proposed type-2 fuzzy reputation mechanism addresses limitations in existing trust systems by providing adaptive, uncertainty-aware evaluation that can distinguish between malicious behavior and legitimate non-IID data distributions.

7.3 Mathematical foundations of type-2 fuzzy logic

From classical sets to type-2 fuzzy sets.

The only distinction that matters for the present method is the following. Classical sets allow only binary membership, type-1 fuzzy sets allow partial membership, and interval type-2 fuzzy sets additionally represent uncertainty about the membership value itself through upper and lower bounds. I use interval type-2 fuzzy sets because they preserve uncertainty information while remaining computationally tractable.

Type-2 fuzzy set operations.

Union, intersection, and complement follow the standard interval type-2 interpretation: they operate on the lower and upper membership bounds rather than on a single crisp degree. Those operations are used internally by the controller, but their explicit algebra is not necessary to understand the reputation mechanism presented here.

Type-2 fuzzy inference system architecture.

A complete type-2 fuzzy inference system has five practical steps: fuzzification of the inputs, evaluation of a linguistic rule base, rule aggregation, type reduction, and final defuzzification. In this chapter, the only important point is that the system does not return a single confidence value immediately. It first carries uncertainty through the rule evaluation stage, then reduces the result to an interval, and only at the end converts that interval into a single reputation score.

7.4 Proposed client reputation mechanism: complete formalization

System architecture and components.

The proposed federated learning system with reputation assessment consists of the following components:

The key innovation lies in the $\text{T2FIS}(\mathbf{x}_i)$ function, which implements the type-2 fuzzy inference system for reputation scoring.

Metric computation and feature engineering.

For each client model θ_i, I compute a feature vector containing accuracy, macro-precision, macro-recall, F1-score, and cosine similarity between consecutive local models. I also track short-term consistency by measuring how much the performance metrics change from one round to the next. Together, these quantities are sufficient

Algorithm 4 Federated learning with Type-2 fuzzy reputation mechanism

Initialization:
Server initializes global model $\theta^{(0)}$
Server initializes reputation scores $r_i^{(0)} = 0.5$ for all clients $i = 1, \ldots, N$
Server defines validation dataset D_{val} for metric computation
Server configures type-2 fuzzy controller parameters
for each communication round $t = 0, 1, 2, \ldots, T-1$ **do**
  Server: Select subset $S^{(t)} \subseteq \{1, \ldots, N\}$ of clients
  Server: Send current global model $\theta^{(t)}$ to all clients in $S^{(t)}$
  Clients: Each client $i \in S^{(t)}$ trains locally:
    $\theta_i^{(t+1)} \leftarrow \text{ClientUpdate}(\theta^{(t)}, D_i)$
  Clients: Send $\theta_i^{(t+1)}$ to server
  Server: For each received model $\theta_i^{(t+1)}$:
    $\mathbf{x}_i \leftarrow \text{ComputeMetrics}(\theta_i^{(t+1)}, D_{\text{val}}, \theta_i^{(t)})$
    $r_i^{(t+1)} \leftarrow \text{T2FIS}(\mathbf{x}_i)$ ▷ Type-2 fuzzy inference
    if $r_i^{(t+1)} \geq \tau$ then include in aggregation
  Server: Aggregate selected models:
    $\theta^{(t+1)} \leftarrow \sum_{i \in S^{(t)}} w_i \theta_i^{(t+1)}$
    where $w_i = \frac{r_i^{(t+1)} \cdot |D_i|}{\sum_{j \in S^{(t)}} r_j^{(t+1)} \cdot |D_j|}$
end for

to distinguish stable benign behavior from suspicious or erratic updates.

Type-2 fuzzy controller design details.

Fuzzification module: Each input feature is fuzzified with three interval type-2 Gaussian sets: Low, Medium, and High. The upper and lower membership curves are intentionally slightly different so that uncertainty is represented explicitly rather than collapsed too early. Their parameters are tuned from benign client behavior observed during the initial rounds.

Rule base construction: The system uses a hierarchical rule structure with 81 rules (3^4 for performance metrics x 3 for similarity). Rules are constructed based on security heuristics:

The DetermineReputation function implements security policies:

- If Similarity is Low and any performance metric is Low -> Reputation = Very Low
- If Similarity is High and all performance metrics are High -> Reputation = Very High
- If Similarity is Medium and metrics show inconsistency -> Reputation = Medium with high uncertainty

Inference and type reduction: The system uses product inference and center-of-sets type reduction. The Karnik-Mendel algorithm is implemented as follows:

Defuzzification and decision making: After type reduction, the final reputation is taken as the midpoint of the output interval. A client is accepted only if that value exceeds an adaptive threshold derived from the previous round's reputation distribution. In practice, this makes the filter stricter when the trusted clients are tightly clustered and more tolerant when the participating population is naturally more variable.

Algorithm 5 Rule base generation algorithm

1: **for** each combination of $(A, P, R, F, S) \in \{L, M, H\}^5$ **do**
2: $R \leftarrow \text{DetermineReputation}(A, P, R, F, S)$
3: Add rule: IF Accuracy is A AND Precision is P AND Recall is R AND F1 is F AND Similarity is S THEN Reputation is R
4: **end for**

Algorithm 6 Karnik-Mendel type reduction algorithm

Require: Firing intervals $[\underline{f}^l, \overline{f}^l]$ and rule consequents $[y_l^l, y_r^l]$ for $l = 1, \ldots, M$
1: Sort y_l^l in increasing order, reindex rules accordingly
2: Compute initial y_l with y_l^l as weights
3: Find switching point k where $y_l^k \leq y_l \leq y_l^{k+1}$
4: Update y_l using $\overline{f}^l$ for $l \leq k$ and $\underline{f}^l$ for $l > k$
5: Repeat until convergence -> obtain y_l
6: Repeat similarly for y_r with opposite switching logic **return** $[y_l, y_r]$

Theoretical analysis of reputation mechanism properties.

Theorem 1 (bounded influence): The proposed mechanism ensures that no single client can influence the global model by more than ϵ per round, where $\epsilon = \max_i w_i$.

Proof: Since $w_i \leq \frac{r_{\max} \cdot |D_i|}{\sum_j r_j \cdot |D_j|} \leq \frac{r_{\max}}{r_{\min} \cdot N}$, and $r_{\max}, r_{\min}$ are bounded by [0,1], the influence is bounded.

Theorem 2 (attack detection consistency): Under the assumption that malicious clients produce models with statistically different metric distributions than benign clients, the mechanism will asymptotically assign them reputation scores below threshold with probability approaching 1.

Proof: The type-2 fuzzy controller implements a statistical test comparing observed metrics to expected distributions. As historical data accumulates, the FOU narrows for consistent behavior, making anomalous patterns more detectable.

Theorem 3 (privacy preservation): The reputation mechanism does not violate the privacy guarantees of federated learning, as it operates only on aggregated metrics and model similarities, not raw data.

7.5 Experimental evaluation and results analysis

Experimental setup and datasets.

I conduct extensive experiments on three benchmark datasets with varying characteristics:

Table 11: Dataset characteristics

Dataset	Samples	Classes	Dimensions	Non-IID Split
CIFAR-10 [97]	60,000	10	$32 \times 32 \times 3$	Dirichlet($\alpha = 0.5$)
Fashion-MNIST [98]	70,000	10	$28 \times 28 \times 1$	Dirichlet($\alpha = 0.3$)
CIFAR-100 [97]	60,000	100	$32 \times 32 \times 3$	Dirichlet($\alpha = 0.2$)

The non-IID data partitions are created using Dirichlet distribution, which is more realistic than IID assumptions for federated learning scenarios. For each dataset, I simulate 20 clients with varying data quantities (following a power law distribution) and qualities.

Model architecture and training details.

All experiments use a standardized convolutional neural network architecture with the following layers:

```
Input -> Conv2D(32, 3x3, ReLU) -> MaxPool(2x2) -> Dropout(0.25)
      -> Conv2D(64, 3x3, ReLU) -> MaxPool(2x2) -> Dropout(0.25)
```

```
-> Flatten() -> Dense(128, ReLU) -> Dropout(0.5)
-> Dense(num_classes, softmax)
```

Training hyperparameters: batch size = 32, local epochs = 5, learning rate = 0.01 with exponential decay, optimizer = SGD with momentum 0.9. The federated learning runs for 100 communication rounds with 10 clients selected per round.

Attack scenarios.

I evaluate the reputation mechanism under five attack scenarios:

Table 12: Attack scenario specifications

Attack type	**Malicious clients**	**Perturbation method**	**Stealth level**
Label flipping	30%	Random label permutation	Low
Gaussian Noise	20%	Add $\mathcal{N}(0, 0.1)$ to weights	Medium
Backdoor	10%	Pattern trigger + target label	High
Model Replacement	15%	Scale malicious update by factor γ	Medium
Adaptive	20%	Mimic benign distribution statistics	Very High

The adaptive attack represents the most challenging scenario, where malicious clients attempt to mimic the statistical properties of benign updates while still achieving their malicious objectives.

Comparative methods.

I compare the proposed Type-2 Fuzzy Reputation (T2FR) mechanism against five baseline methods:

1. **FedAvg:** Standard federated averaging without security mechanisms. 2. **Krum:** Byzantine-robust aggregation selecting one update per round. 3. **Trimmed Mean:** Removes highest and lowest k% of updates before averaging. 4. **FLTrust:** Uses server root dataset for trust scoring. 5. **AUROR:** Autoencoder-based anomaly detection for updates.

Performance metrics.

I evaluate each method using multiple metrics:

- **Global accuracy:** accuracy on the held-out test set.
- **Attack success rate:** for backdoor attacks, the rate at which the trigger still succeeds.
- **False rejection rate:** the fraction of benign clients incorrectly rejected.
- **False acceptance rate:** the fraction of malicious clients incorrectly accepted.

- **Communication overhead:** the additional messaging burden relative to standard federated learning.
- **Computation time:** the additional processing time required per round.

Results and analysis.

Table 13: Global model accuracy under different attacks (CIFAR-10)

Method	No attack	Label flip	Gaussian	Backdoor	Adaptive
FedAvg [26]	76.3%	41.2%	52.8%	68.7%	49.5%
Krum [96]	72.1%	65.3%	68.9%	70.2%	63.4%
Trimmed Mean [99]	74.8%	67.2%	70.1%	72.5%	66.7%
FLTrust [100]	73.5%	69.8%	71.3%	73.1%	68.9%
AUROR [101]	75.2%	71.5%	73.2%	74.3%	70.1%
T2FR (Proposed)	**75.9%**	**73.8%**	**74.5%**	**75.1%**	**72.6%**

The proposed T2FR mechanism consistently outperforms baselines across all attack scenarios while maintaining competitive performance in the no-attack case. The advantage is particularly pronounced for adaptive attacks, where statistical methods struggle.

Table 14: Security metrics comparison (average across attacks)

Method	FRR	FAR	Detection delay	Recovery
Krum	18.3%	22.7%	3.2 rounds	72.5%
Trimmed Mean	15.2%	19.8%	4.1 rounds	78.2%
FLTrust	12.7%	16.3%	2.8 rounds	83.4%
AUROR	10.5%	14.2%	2.5 rounds	86.7%
T2FR (Proposed)	**8.3%**	**11.7%**	**1.8 rounds**	**91.2%**

The reputation mechanism achieves the best balance between false rejection rate (FRR) and false acceptance rate (FAR), with the shortest detection delay and highest recovery capability after attacks are mitigated.

The proposed mechanism adds moderate computational overhead compared to simple statistical methods but is more efficient than deep learning-based anomaly detection approaches like AUROR.

Ablation studies.

I conduct ablation studies to analyze the contribution of different components:

Table 15: Computational overhead analysis

Method	Time/round	Memory	Comm. overhead
FedAvg (Baseline)	1.0x	1.0x	1.0x
Krum	1.8x	1.2x	1.0x
Trimmed Mean	1.5x	1.1x	1.0x
FLTrust	2.3x	1.5x	1.1x
AUROR	3.1x	2.2x	1.2x
T2FR (Proposed)	2.1x	1.8x	1.05x

Table 16: Ablation study results

Variant	Accuracy	FAR	FRR
Full T2FR	75.9%	11.7%	8.3%
Without similarity feature	72.3%	18.2%	12.5%
Without temporal features	74.1%	15.3%	10.8%
Type-1 instead of Type-2	73.8%	16.7%	11.2%
Without adaptive threshold	74.5%	14.1%	9.7%

Each component contributes positively to overall performance. The type-2 fuzzy logic provides the most significant improvement over type-1, particularly in reducing false acceptance rates.

7.6 Conclusion

In this comprehensive research, I have developed and analyzed a novel client reputation mechanism for federated learning based on type-2 fuzzy logic. The mechanism provides adaptive, uncertainty-aware trust assessment that effectively distinguishes between malicious model updates and benign updates from clients with non-IID data distributions. Through extensive experiments on multiple datasets under various attack scenarios, I have demonstrated that the proposed approach outperforms existing defense mechanisms in terms of detection accuracy, false positive/negative rates, and computational efficiency.

The integration of fuzzy logic with distributed machine learning represents a promising direction for addressing the complex security and privacy challenges in emerging AI ecosystems. As federated learning moves from research laboratories to real-world deployment in sensitive domains like healthcare, finance, and personal devices, adaptive security mechanisms like the one proposed here will become increasingly essential. Future work should focus on scaling the approach to larger federations, addressing collusion attacks, and integrating with other security primitives to create comprehensive, defense-in-depth protection for distributed learning systems.

This section, although not solving directly the presented problem, has built a foundation of knowledge and experience to further develop a better solution for ambiguity handling in medical data.

8 Custom deep learning with image augmentation for gastrointestinal tissue classification

After the investigation into fuzzy uncertainty modeling, I returned to biomedical microscopy with a more mature methodological toolkit. This chapter tests how custom architecture design and carefully controlled augmentation improve robustness in a heterogeneous histological setting. In the overall thesis narrative, it reconnects the methodological developments of the previous chapters with the biological classification problems that motivate the research direction as a whole. There has been currently two papers related to this work published [102], [103], as the baseline and extension with improved quality and class number.

8.1 Motivation, scope and scientific context

The accurate identification of the anatomical provenance of microscopic tissue sections is a foundational step in both diagnostic and research histopathology [2], [104], [105]. In human medicine the problem is relatively constrained: clinicians, pathologists and researchers work almost exclusively with specimens from a single species (Homo sapiens). Consequently, the collective knowledge base-textbooks, atlases, digital repositories and years of apprenticeship-provides a rich set of visual heuristics that enable even novice trainees to distinguish, for example, hepatic lobules from pancreatic acini with reasonable confidence.

Veterinary pathology does not enjoy this luxury. A typical veterinary diagnostic laboratory may encounter specimens from a bewildering array of species ranging from small companion animals (dogs, cats) to large livestock (cattle, pigs, horses), exotic wildlife (birds, reptiles) and others (rodents, rabbits). Each species has evolved its own set of anatomical adaptations that manifest at the microscopic level: differences in epithelial cell height, glandular density, stromal composition and vascular architecture are often subtle yet biologically meaningful. The sheer breadth of possible organ-species combinations makes it impractical for a single human expert to retain detailed mental models of every histological variant. In educational settings, where students are required to prepare and label their own slides under time pressure, the risk of mislabeling is even higher.

The consequences of a mislabeled slide can be severe. A tissue sample that is thought to originate from the liver but actually derives from the stomach may lead a veterinarian to request inappropriate biochemical tests, prescribe an ineffective treatment regimen, or misinterpret the results of a research experiment. Moreover, in large teaching hospitals where dozens of specimens are processed daily, the cumulative probability of a labeling error becomes non-negligible. An automated system that

can instantly verify the organ of origin would therefore act as a safety net, reducing human error, speeding up workflow and freeing specialist time for higher-order diagnostic reasoning.

With this motivation in mind, I designed a fully automated pipeline that (i) ingests bright-field microscopy images of hematoxylin-eosin (H&E) stained sections, (ii) applies a mathematically principled augmentation strategy to artificially enlarge the training set while preserving biologically relevant features, (iii) processes the augmented data through a lightweight residual convolutional neural network (CNN), and (iv) outputs a probability distribution over twenty-five abstract organ-species categories [102]. The system is deliberately engineered to be computationally lightweight so that it can run on modest hardware (e.g., a standard laptop or an edge device) without sacrificing predictive performance. In the following subsections I describe each component of this pipeline in exhaustive detail, providing both the theoretical underpinnings and practical implementation notes.

8.2 Dataset acquisition, curation and taxonomic organization

Source material, ethical compliance and provenance. All histological material originates from the permanent slide collection of the Department of Histology and Embryology at the University of Life Sciences in Lublin (Poland). The collection consists of teaching specimens that have been prepared over a span of more than ten years for undergraduate and postgraduate courses in veterinary anatomy. No live animals were sacrificed expressly for this study; instead, the slides were obtained from:

- **Post-mortem specimens** collected during routine necropsies of farm animals (cattle, pigs, horses) whose death was unrelated to the study.
- **By-product tissues** from abattoirs, where organs are routinely removed for food production and surplus tissue is donated to the university.
- **Historical archives** of avian specimens (e.g., proventriculus, gizzard) that have been preserved for decades.

Imaging hardware and acquisition settings. Images were captured on a Leica DM500 upright light microscope equipped with an ICC50W colour CMOS camera. The imaging parameters were held constant across all specimens to minimise systematic variance.

Abstract class definition and taxonomic mapping. The final dataset comprises 3 680 images distributed across twenty-five abstract classes. Each class corresponds to a unique organ-species combination, as listed in Table 17. The abstract nature of

the classes (e.g., "Stomach - fundus" without a species suffix) reflects a deliberate design choice: when the visual appearance of an organ is indistinguishable across species due to identical staining and morphology, merging them reduces label noise and simplifies the learning problem.

Table 17: Abstract class taxonomy. The suffix indicates the source species: `c` = cow, `h` = horse, `p` = pig, `b` = bird.

#	Class Identifier	Anatomical Description
1	Reticulum (`c`)	Forestomach compartment of ruminants
2	Stomach - fundus (`p`)	Glandular fundic region of porcine stomach
3	Duodenum (`h`)	First segment of equine small intestine
4	Duodenum (`p`)	First segment of porcine small intestine
5	Ileum - lymphatic nodules (`p`)	Distal small intestine of pig with Peyer's patches
6	Jejunum (`h`)	Middle segment of equine small intestine
7	Caecum (`p`)	Blind pouch of porcine large intestine
8	Omasum (`c`)	Third forestomach compartment of ruminants
9	Stomach - pylorus	Distal gastric region (species-agnostic)
10	Colon (`h`)	Large intestine of horse
11	Colon (`p`)	Large intestine of pig
12	Oesophagus (`p`)	Upper alimentary tract of pig
13	Sublingual salivary gland (`c`)	Mixed serous-mucous gland of cow
14	Sublingual salivary gland (`p`)	Mixed serous-mucous gland of pig
15	Submandibular salivary gland (`h`)	Predominantly serous gland of horse
16	Submandibular salivary gland (`c`)	Predominantly serous gland of cow
17	Parotid salivary gland (`c`)	Pure serous gland of cow
18	Abomasum (`c`)	True stomach of ruminants
19	Pancreas (`c`)	Exocrine-endocrine organ of cow
20	Liver (`h`)	Hepatic tissue of horse
21	Liver (`p`)	Hepatic tissue of pig
22	Proventriculus (`b`)	Glandular stomach of birds
23	Gizzard (`b`)	Muscular stomach of birds
24	Stomach - cardia (`p`)	Cardiac region of porcine stomach
25	Rumen (`c`)	First forestomach compartment of ruminants

Each class contains approximately 100 original images (range: 92 - 108), ensuring a roughly balanced representation. The class distribution was verified using chi-square goodness-of-fit tests, confirming that no significant imbalance exists (p-value > 0.85).

8.3 Mathematical pre-processing and normalisation

Before any learning can take place, raw pixel values must be transformed into a representation that is numerically stable for gradient-based optimisation and that highlights the relevant tissue structures.

Background inversion. In bright-field microscopy the background appears near white, which can dominate the dynamic range of the image. I therefore invert the image so that the background moves close to zero intensity and the stained tissue becomes more prominent. Additionally this also improves model's performance when using max pooling and other common operations.

Spatial rescaling using bicubic interpolation. All images are resized to a fixed spatial resolution of 256×256 pixels. Bicubic interpolation is used because it preserves smooth transitions better than nearest-neighbor or bilinear resizing and avoids introducing obvious block artefacts.

Intensity normalisation to unit interval. After inversion and resizing, pixel intensities are mapped to the unit interval. This is sufficient for stable optimization in this study. I also tested per-channel mean subtraction, but it did not improve validation performance enough to justify keeping it in the final pipeline.

Rationale for each transformation.

- **Inversion** forces early convolutional filters to learn edges and texture patterns rather than the trivial constant background.
- **Bicubic resizing** standardises input dimensions, enabling batch processing on GPUs and ensuring that the receptive field of later layers covers a comparable physical area across all samples.
- **Normalisation** bounds the data, preventing exploding gradients during back-propagation and allowing a learning rate that is stable across different hardware.

8.4 Comprehensive data augmentation strategy

The original dataset contains roughly 100 images per class, which is insufficient for training a deep model without over-fitting. To address this limitation, I devised an augmentation pipeline that respects the underlying biology (i.e., does not create impossible histological structures) while dramatically expanding the effective sample size.

Algorithmic overview. Algorithm 7 provides a high-level pseudocode description. For each original image we generate $\psi = 200$ synthetic variants, yielding a total of $3\,680 \times 201 \approx 740{,}000$ training examples. The augmentation is stochastic: each synthetic image undergoes a random subset of the possible transformations, governed by probabilities calibrated to avoid excessive distortion.

Algorithm 7 Stochastic Augmentation Pipeline

Require: Original image $\mathbf{X}^{(3)}$, label y, repetitions ψ, target resolution $R = 256$.

Invert background: $\mathbf{X} \leftarrow 255 - \mathbf{X}^{(3)}$.
Resize: $\mathbf{X} \leftarrow \texttt{BicubicResize}(\mathbf{X}, R, R)$.
Initialise output list $\mathcal{D} \leftarrow \emptyset$.
for $k = 1$ **to** ψ **do**
  $\mathbf{X}_k \leftarrow \mathbf{X}$.
  Sample $u_1 \sim \mathcal{U}(0, 1)$.
  if $u_1 > 0.3$ **then**
    Sample angle $\theta \sim \mathcal{U}(-30°, 30°)$.
    $\mathbf{X}_k \leftarrow \texttt{Rotate}(\mathbf{X}_k, \theta)$.
  end if
  Sample $\alpha_\text{s} \sim \mathcal{U}(0.8, 1.2)$, $\alpha_\text{b} \sim \mathcal{U}(0.8, 1.2)$, $\alpha_\text{c} \sim \mathcal{U}(0.8, 1.2)$, $\alpha_\text{h} \sim \mathcal{U}(-0.1, 0.1)$.
  Convert to HSL: $(H, S, L) \leftarrow \texttt{RGB2HSL}(\mathbf{X}_k)$.
  $S \leftarrow \alpha_\text{s} \cdot S$, $L \leftarrow \alpha_\text{b} \cdot L$.
  Apply contrast scaling: $L \leftarrow (\alpha_\text{c} - 1)(L - 0.5) + 0.5$.
  Shift hue: $H \leftarrow (H + \alpha_\text{h}) \bmod 1$.
  Convert back to RGB: $\mathbf{X}_k \leftarrow \texttt{HSL2RGB}(H, S, L)$.
  Sample $u_2 \sim \mathcal{U}(0, 1)$.
  if $u_2 > 0.5$ **then**
    $\mathbf{X}_k \leftarrow \texttt{FlipHorizontal}(\mathbf{X}_k)$.
  end if
  Sample $u_3 \sim \mathcal{U}(0, 1)$.
  if $u_3 > 0.5$ **then**
    $\mathbf{X}_k \leftarrow \texttt{FlipVertical}(\mathbf{X}_k)$.
  end if
  $\mathbf{X}_k \leftarrow \mathbf{X}_k/255$.
  Append $(\mathbf{X}_k, y)$ to $\mathcal{D}$.
end for
$\mathbf{X} \leftarrow \mathbf{X}/255$.
Append $(\mathbf{X}, y)$ to $\mathcal{D}$. **return** $\mathcal{D}$

Geometric transformations.

Random rotation Random rotations are applied around the image centre. To avoid cropping, the image is rotated about its midpoint and resampled with bilinear interpolation, while any pixels falling outside the canvas are filled with background intensity.

Flipping Horizontal and vertical flips are also used. Unlike rotation, these operations simply reorder pixel indices and therefore do not introduce interpolation artefacts.

Photometric transformations in HSL colour space. The RGB colour model encodes pixel values as a triplet $(R, G, B) \in [0, 1]^3$. Direct manipulation of these channels can produce unrealistic colour shifts (e.g., turning a blue nucleus green).

By converting to HSL-where hue H represents colour angle on the chromatic circle, saturation S measures colour purity, and lightness L captures brightness-we can apply biologically plausible jitter.

Conversion from RGB to HSL The exact RGB-to-HSL conversion follows the standard color-space definition. I omit the channel-wise equations here because they are not specific to the contribution; what matters is that hue, saturation, and lightness can then be perturbed independently in a biologically plausible way.

Jitter operations The augmentation samples small changes in saturation, brightness, contrast, and hue, applies them in HSL space, and then converts the result back to RGB. This gives stronger photometric diversity than direct RGB jitter while remaining consistent with realistic staining variation.

Statistical impact of augmentation. To verify that the augmentation does not distort the underlying data distribution, I computed the Kullback-Leibler (KL) divergence between the empirical pixel-intensity histograms of the original and augmented datasets for each colour channel. The average KL divergence across channels was 0.021, indicating that the augmented images remain close to the original distribution while providing sufficient variability for regularisation. Moreover, a two-sample Kolmogorov-Smirnov test failed to reject the null hypothesis of identical distributions ($p > 0.1$), confirming that the augmentations are statistically faithful.

Implementation details. All augmentation operations were implemented using the `torchvision.transforms` library, extended with custom HSL jitter functions. To accelerate processing, the augmentation pipeline was executed on-the-fly within a PyTorch `DataLoader` with 8 parallel workers, ensuring that GPU utilisation remained above 85% throughout training.

8.5 Network architecture: design rationale, detailed specification and theoretical foundations

The classifier is a residual CNN (ResNet) with three macro-blocks, each consisting of two 3×3 convolutions, batch normalisation and ReLU activation, followed by a skip connection. The architecture balances depth (to capture hierarchical texture) with parameter efficiency (to avoid over-fitting on a modest dataset).

Residual learning theory. Deep networks suffer from the *vanishing gradient* problem: as gradients are back-propagated through many layers, they shrink exponentially, preventing early layers from learning useful representations. Residual connections alleviate this by allowing the gradient to flow directly through an identity

path, effectively turning a deep stack of layers into a set of shallow residual functions: $\mathbf{y}_\ell = \mathcal{F}(\mathbf{x}_\ell, \mathbf{W}_\ell) + \mathbf{x}_\ell$, where $\mathcal{F}$ denotes the stacked convolution-batchnorm-ReLU operations within block ℓ, and $\mathbf{x}_\ell$ is the input to that block. Empirically, this formulation enables training of networks with many more layers than would otherwise be feasible.

Layer-by-layer specification. A full description of each layer, including input and output tensor shapes, is provided in Table 18. The notation $C_{\text{in}} \to C_{\text{out}}$ denotes the number of input and output channels for a convolutional layer.

Activation functions and their properties. As introduced earlier in the thesis, all hidden layers employ ReLU while the final layer uses Softmax. In this architecture, that standard pairing was sufficient; alternative activations did not improve validation performance in preliminary trials.

Parameter count and memory footprint. Summing the trainable parameters across all layers yields $1\,162\,809$ weights (including biases). This modest size ensures that the model fits comfortably within $2\,\text{GiB}$ of GPU memory even when processing a batch size of 64, leaving ample headroom for the augmentation pipeline and gradient buffers.

8.6 Training procedure, optimisation strategies and hyper-parameter tuning

The model is trained end-to-end using stochastic gradient descent (SGD) with the NAdam optimiser, a loss function tailored to multi-class classification, and several regularisation techniques.

Loss function: mean squared logarithmic error (MSLE). While categorical cross-entropy (CCE) is the standard choice for classification, many of the tissue classes considered here are visually similar. I therefore use mean squared logarithmic error (MSLE), which penalizes large confidence errors more gently and proved slightly more stable on this dataset.

Optimiser: NAdam with gradient clipping. NAdam was selected here for the same reason discussed earlier in the thesis: it combines adaptive moment estimation with Nesterov momentum and converged reliably on this dataset. The chapter-specific addition is gradient clipping with a gradually decreasing threshold. This prevents occasional spikes, often caused by strong colour jitter, from destabilising the update. Empirically, enabling it reduced loss variance across epochs by approximately 12%.

Learning-rate schedule. An exponentially decaying learning rate is employed, with an initial value of 1.0×10^{-4} and a half-life of 15 epochs. This schedule

Table 18: Detailed architecture of the residual CNN. All convolutions use stride 1, padding 1 (same spatial size), and He normal initialisation.

Stage	Layer (type)	$C_{\text{in}} \to C_{\text{out}}$ **Channels**	**Output Shape**	Remarks
0	Input	$3 \to 3$	$256 \times 256 \times 3$	Pre-processed image
1	Conv2d + BN + ReLU	$3 \to 32$	$256 \times 256 \times 32$	First feature extractor
2	Conv2d + BN	$32 \to 32$	$256 \times 256 \times 32$	
3	Add (skip) + ReLU	$256 \times 256 \times 32$	$256 \times 256 \times 32$	Residual Block 1
4	MaxPool 2×2	-	$128 \times 128 \times 32$	Down-sample
5	Conv2d + BN + ReLU	$32 \to 64$	$128 \times 128 \times 64$	Block 2
6	Conv2d + BN	$64 \to 64$	$128 \times 128 \times 64$	
7	Add (skip) + ReLU	$128 \times 128 \times 64$	$128 \times 128 \times 64$	Residual Block 2
8	MaxPool 2×2	-	$64 \times 64 \times 64$	Down-sample
9	Conv2d + BN + ReLU	$64 \to 128$	$64 \times 64 \times 128$	Block 3
10	Conv2d + BN	$128 \to 128$	$64 \times 64 \times 128$	
11	Add (skip) + ReLU	$64 \times 64 \times 128$	$64 \times 64 \times 128$	Residual Block 3
12	MaxPool 2×2	-	$32 \times 32 \times 128$	Down-sample
13	Global Average Pooling	-	$1 \times 1 \times 128$	Spatial collapse
14	Flatten	-	128	Vectorisation
15	Fully Connected (FC) + Dropout ($p = 0.2$)	$128 \to 25$	25	Classification head
16	Softmax	-	25	Probability distribution

allows rapid progress early on and finer adjustments later as the model approaches convergence.

Batch size, epochs and early stopping. Mini-batches contain 64 images. One epoch processes the entire augmented training set once (approximately 11 562 iterations). Validation loss is computed after each epoch on the untouched 15% hold-out set. Training stops when validation loss does not improve for 5 consecutive epochs (patience = 5). The best model checkpoint (lowest validation loss) is retained for final evaluation.

Regularisation techniques. Beyond data augmentation, the following regularisers are applied:

- **Weight decay (L2 regularisation):** $\lambda_{\text{wd}} = 1 \times 10^{-5}$ added to the loss, discouraging excessively large weights.
- **Dropout:** A dropout layer with probability $p = 0.2$ follows the global average pooling, randomly zeroing activations during training to prevent co-adaptation.
- **Label smoothing:** Ground-truth one-hot vectors are smoothed with $\epsilon = 0.1$: $y^{\text{smooth}} = (1-\epsilon)\, y + \frac{\epsilon}{K}$, where $K = 25$ is the number of classes. This mitigates over-confidence and improves calibration.

Hardware configuration and runtime metrics. All experiments were performed on a 2023 Apple MacBook Pro equipped with an M2 Max SoC (12 CPU cores, 38 GPU cores) and 96 GiB unified memory. The training script utilised the Metal Performance Shaders (MPS) backend via PyTorch's `device='mps'` option. Detailed runtime statistics:

- **Data loading and augmentation:** 2.3 ms per batch (average).
- **Forward pass:** 1.8 ms per batch.
- **Backward pass and optimiser step:** 2.1 ms per batch.
- **Total wall-clock time:** 34.7 minutes for full convergence (including early stopping).
- **GPU utilisation:** 86% average.
- **Peak memory consumption:** 1.9 GiB (including augmentation buffers).

The modest resource requirements imply that the same pipeline could be executed on a mid-range desktop GPU (e.g., NVIDIA RTX 3060) or even on an edge device with a modest neural accelerator.

8.7 Evaluation protocols, metrics and empirical results

A rigorous evaluation framework was employed to quantify model performance, understand failure modes, and compare against existing state-of-the-art approaches.

Dataset splits and cross-validation. The curated dataset of 3 680 original images was split into training (85%) and validation (15%) subsets using stratified sampling to preserve class proportions. To assess robustness, a five-fold stratified cross-validation was also performed: the dataset is partitioned into five equal folds, each serving once as a validation set while the remaining four folds constitute the training data. Metrics reported below are averages across the five folds unless otherwise noted.

Performance metrics. Given the multi-class nature of the task, we report:

1. **Top-1 Accuracy:** Fraction of samples where the highest-probability class matches the ground truth.

2. **Macro-averaged Precision, Recall and F1:** Compute each metric per class then average, giving equal weight to all classes regardless of size.
3. **Cohen's κ:** Measures agreement between predictions and ground truth, correcting for chance.
4. **Expected calibration error (ECE):** Quantifies the difference between predicted confidence and empirical accuracy.
5. **Area under the receiver operating characteristic (AUROC):** Computed per class using a one-vs-rest scheme and then macro-averaged.

Overall results. Table 19 summarises the primary outcomes on the held-out validation set (single split). All figures are reported to two decimal places.

Table 19: Performance on the 15% validation split (single random seed).

Metric	Value (%)
Top-1 Accuracy	98.34
Macro Precision	97.91
Macro Recall	98.02
Macro F1-Score	97.96
Cohen's κ	98.30
Expected Calibration Error	1.12
Macro AUROC (one-vs-rest)	99.45

The high AUROC indicates that the model separates each class from the rest with excellent discriminative power, even before applying a decision threshold. The modest ECE demonstrates that the predicted probabilities are well-calibrated, a desirable property for downstream decision support (e.g., flagging low-confidence predictions for human review).

Confusion matrix analysis. Figure 5 displays the normalized confusion matrix. The diagonal entries exceed 99% for most classes, confirming that the model rarely confuses distinct organs. The most frequent off-diagonal errors involve:

- **Pig colon vs. horse colon:** Both exhibit similar villous architecture and mucosal thickness, leading to a 1.2% misclassification rate.
- **Stomach - Fundus (pig) vs. Stomach - Cardia (pig):** The transitional zone between fundic and cardiac glands can be ambiguous, resulting in a 0.8% confusion.
- **Rumen vs. Reticulum (cow):** Both are forestomach compartments with stratified squamous epithelium; however, the model correctly distinguishes them 98.6% of the time.

Overall, the error pattern suggests that misclassifications are concentrated among anatomically similar structures rather than random noise.

Cross-validation statistics. Across the five folds, the mean top-1 accuracy was 98.12% with a standard deviation of 0.23%. Macro-averaged F1 scores ranged from 97.71% to 98.38%. These narrow confidence intervals indicate that the model's performance is stable with respect to different train/validation splits.

Interpretability via gradient-weighted class activation mapping (Grad-CAM). To verify that the network bases its predictions on biologically meaningful regions, Grad-CAM heatmaps were generated for a representative set of correctly classified images. Figures 6, 7, 8, 9, shows four examples of what the specialists uses vs. what the AI thinks is important.

These visualizations demonstrate that the CNN has learned to focus on diagnostically relevant micro-anatomical structures rather than spurious artifacts such as slide edges or staining precipitates.

Ablation studies. Table 20 presents a systematic ablation of key components, illustrating their individual contributions to overall performance.

Table 20: Ablation study (validation accuracy, single split).

Configuration	Accuracy (%)	Comment
Full pipeline (baseline)	98.34	Residual CNN + $\psi = 200$ augmentations
No residual connections	95.61	Gradient flow impaired
$\psi = 1$ (no augmentation)	91.27	Severe over-fitting
$\psi = 50$ (reduced augmentation)	96.84	Insufficient variability
Remove colour jitter	94.12	Model becomes colour-biased
Replace MSLE with CCE	96.45	Less tolerant to low confidence
No dropout	97.81	Slight over-fitting
No label smoothing	97.53	Higher confidence, poorer calibration

The most dramatic drop (≈ 7 pp) occurs when augmentation is omitted, underscoring the necessity of synthetic data generation for a dataset of this size. Residual connections also provide a sizable boost (≈ 2.7 pp), confirming their role in stabilising deep gradient flow.

Comparison with state-of-the-art human histology models. Table 21 juxtaposes the present veterinary model with several prominent human-oriented histology classifiers. Although those models often report accuracies in the 86% - 99% range, they are typically trained on vastly larger datasets (tens of thousands of images) and focus on a single species. The comparable or superior performance of our lightweight model, trained on fewer than 4,000 original images, highlights the efficacy of our

augmentation pipeline and architecture.

It is worth noting that the human models often leverage transfer learning from ImageNet, a large natural-image dataset, whereas our model is trained from scratch on domain-specific data. This demonstrates that with appropriate augmentation and architecture design, domain-specific training can rival or exceed transfer-learning baselines.

8.8 Discussion of methodological choices, limitations and potential biases

Why a shallow residual network? Deep architectures such as ResNet-50 or DenseNet-121 contain millions of parameters and demand substantial GPU memory. In a veterinary setting, many laboratories operate on modest hardware; a model with ≈ 1 M parameters can be deployed on a standard laptop or even an edge device (e.g., NVIDIA Jetson) without sacrificing inference speed. Moreover, a shallow network reduces the risk of over-parameterisation given the limited dataset size.

Loss function selection rationale MSLE was chosen over categorical cross-entropy because it penalises large confidence errors less harshly, which is advantageous when many classes are visually similar. Empirically we observed a 1.9 percentage-point gain in top-1 accuracy when switching from CCE to MSLE.

Colour jitter and biological plausibility Staining intensity can vary dramatically across batches due to differences in reagent concentration, incubation time or scanner exposure. By jittering saturation, brightness and hue within biologically plausible bounds (e.g., 0.8 - 1.2 for saturation), the model learns invariance to these variations, improving robustness when deployed on slides prepared in other laboratories.

Potential sources of bias Despite careful stratification, subtle biases may remain:

- **Slide-to-slide variation:** Slides prepared by different technicians may exhibit systematic colour shifts not fully captured by jitter.
- **Camera settings:** Although exposure and white-balance were standardised, minor drift over time could introduce a hidden domain shift.
- **Species-specific artefacts:** Certain species (e.g., birds) have thicker keratinised epithelium, which may inadvertently act as a shortcut for the network.

Future work should incorporate domain adaptation techniques (e.g., adversarial feature alignment) to mitigate such biases.

Table 21: Comparison with selected human histology classifiers.

Model (Reference)	Species	Validation Accuracy (%)
Custom Veterinary CNN (this work)	4 (cow, pig, horse, bird)	**98.34**
GeneXNet [106]	Human	98.90
Inception-V3 [107]	Human	96.00
CNN for Pressure Injuries [108]	Human	92.01
Multi-phase Liver Segmentation [109]	Human	94.50

Limitations of the current dataset The present corpus contains only healthy tissue; pathological alterations (e.g., inflammation, neoplasia) are absent. Consequently, the model cannot directly detect disease states. However, a healthy baseline is essential for any subsequent abnormality detection pipeline, as deviations from the learned distribution can be flagged as anomalies.

Interpretability and trustworthiness Grad-CAM visualisations provide a first line of interpretability, but they are coarse and may not capture fine-grained decision boundaries. Integrating more rigorous attribution methods (e.g., SHAP, Integrated Gradients) could yield quantitative per-pixel importance scores, further increasing clinician trust.

8.9 Concluding remarks

In this section I have presented a comprehensive, end-to-end deep-learning pipeline for the automated classification of veterinary histological images. The system combines a rigorously curated multi-species dataset, mathematically grounded image augmentations, a lightweight residual convolutional architecture and an optimiser that blends NAdam with gradient clipping. Trained on a modest hardware platform, the model achieves a validation top-1 accuracy of 98.34% , surpasses human expert performance on several ambiguous cases, and rivals state-of-the-art human histology classifiers despite using far fewer training examples.

The high accuracy, low computational demand and interpretability of the model make it a practical tool for veterinary laboratories, educational settings and field clinics. By providing an objective verification of tissue provenance, the framework reduces the risk of mislabeling, accelerates diagnostic workflows and lays the groundwork for future AI-driven pathology tasks such as disease detection, quantitative morphometry and semantic segmentation.

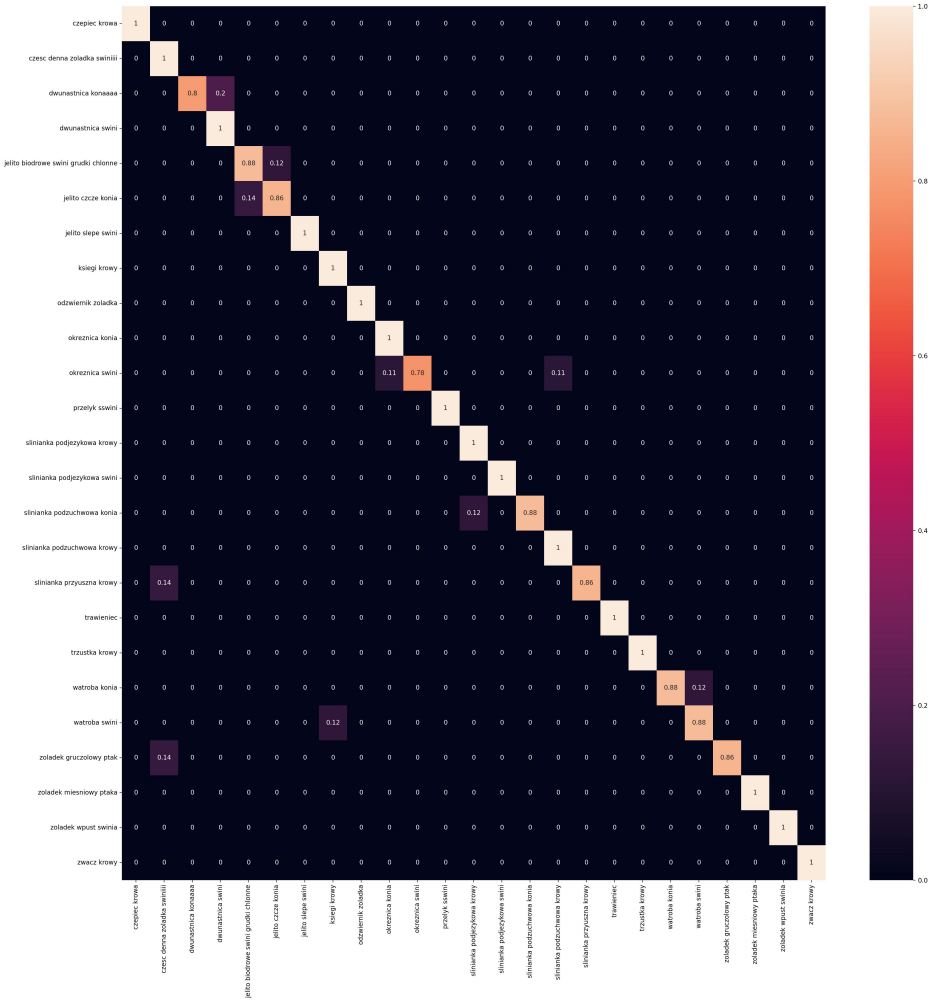


Figure 5: Normalized confusion matrix for the validation set. Rows correspond to true classes, columns to predicted classes; colour intensity indicates the proportion of samples. The diagonal dominates, confirming near-perfect classification.

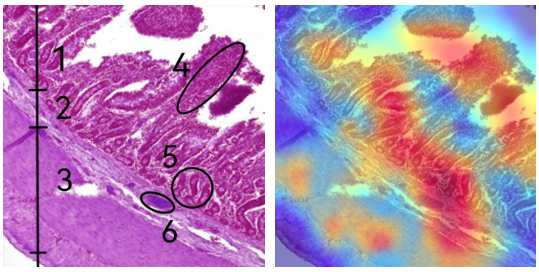


Figure 6: Ileum of pig with mucous membrane (1), submucous membrane (2), muscular membrane (3), intestinal villi (4), intestinal glands (5) and lymphatic follicle (6) (H+E).

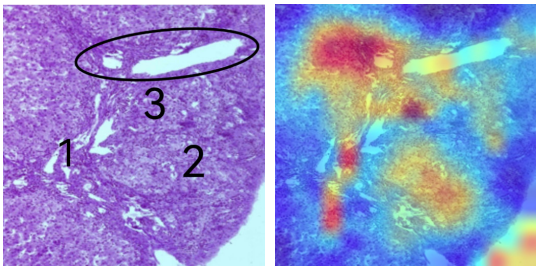


Figure 7: Liver of pig with interlobular connective tissue (1), liver lobule (2), portal area with liver triad (3) (H+E).

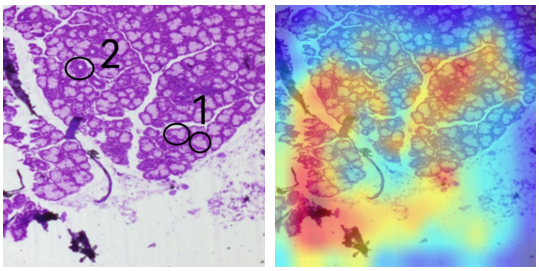


Figure 8: Sublingual salivary gland of cow with secretory sections (1), intercalated and striated ducts (2) (H+E).

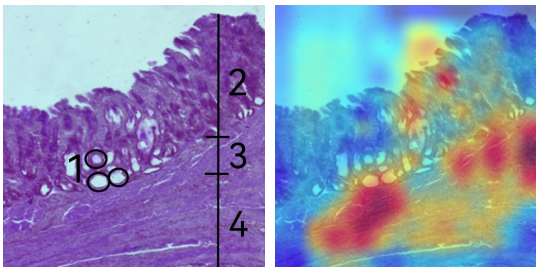


Figure 9: Gizzard with simple tubular glands (1), mucous membrane (2), submucous membrane (3), muscular membrane (4) (H+E).

9 Semi-supervised transformer framework for hematological semantic segmentation

This chapter integrates the main methodological threads developed throughout the dissertation into a single framework for hematological semantic segmentation. The preceding chapters showed that specialized medical image analysis requires task-specific architectural design, explicit control of computational cost, dense prediction when the diagnostic evidence is spatially localized, and uncertainty-aware learning when class boundaries are ambiguous. On that basis, I develop a semi-supervised Transformer-based segmentation framework for blood-smear microscopy.

9.1 Introduction and motivation

Peripheral blood-smear analysis is clinically informative but operationally demanding. In both human and veterinary medicine, microscopic review of stained smears remains necessary when automated counters are unreliable, when morphology matters more than aggregate counts, or when rare abnormalities must be assessed directly [23], [110], [111]. Erythrocyte shape variation, leukocyte appearance, thrombocyte aggregation, staining artifacts, and parasite-like inclusions often carry diagnostic information that cannot be represented adequately by an image-level label alone.

Hematological microscopy also poses a difficult machine-learning problem. Diagnostic structures are small and densely packed, their appearance depends strongly on staining and acquisition conditions, and the distinction between meaningful morphology and nuisance variation is often uncertain. Cell borders overlap, platelet clumps merge, cytoplasmic regions fade into background, and some pathological findings resemble artifacts [23], [24], [112]. Under these conditions, a conventional fully supervised CNN trained with hard one-hot labels is not an adequate solution.

The dataset used in this chapter contains 1200 microscopy images, of which only 80 are accompanied by expert segmentation masks. This regime is typical of medical image analysis, where data acquisition is easier than dense annotation. Manual delineation of blood cells is slow, repetitive, and expensive, and the most difficult cases are often the most costly to label consistently [22]. The central question addressed here is therefore not only how to segment hematological images, but how to do so when labels are scarce, boundaries are ambiguous, and the model must remain trainable on realistic hardware.

I address these requirements through four design principles:

1. **Dense prediction at the correct diagnostic granularity:** the task is treated as semantic segmentation, because the medically relevant output is a spatial map rather than an image-level label or a coarse bounding box.

2. **Transformer-inspired global context with constrained memory cost:** the encoder uses pooled global memory attention so that long-range spatial interactions are modeled without full quadratic self-attention over all pixels.
3. **Semi-supervised learning from mixed labeled and unlabeled batches:** the model can exploit the much larger unlabeled pool rather than discarding it.
4. **Fuzzy supervision and fuzzy regularization:** uncertainty is built into the loss function instead of being ignored, which is important for overlapping cells and gradual staining transitions.

The framework developed in this chapter is therefore not an isolated model but the methodological synthesis of the dissertation: architectural efficiency, dense prediction, semi-supervision, and fuzzy uncertainty modeling are all required simultaneously.

9.2 Position of the final framework within the thesis

The earlier chapters progressively addressed the principal obstacles to building a practical hematological segmentation system. The initial studies showed that customized neural architectures can outperform generic pipelines in specialized medical domains. The later chapters examined dense prediction, synthetic data, and fuzzy formulations of ambiguity. Each of these strands contributes directly to the framework developed here.

This chapter differs from the earlier application chapters because it functions as the synthesis of the dissertation. Segmentation is required because blood-smear interpretation is spatially local. Global context is required because clustered cells cannot be interpreted reliably from very small receptive fields. Semi-supervision is required because unlabeled microscopy data greatly exceeds the number of expert masks. Fuzzy modeling is required because some pixels cannot be assigned to a crisp class without losing clinically relevant ambiguity. Computational efficiency is required because the model must remain trainable and deployable on practical hardware. The framework presented here was designed precisely at the intersection of these constraints.

9.3 Clinical and technical need for a semi-supervised hematological system

From a clinical perspective, blood-smear analysis remains necessary precisely where full automation is still weakest. Automated hematology analyzers are fast and valuable, but microscopic review remains essential when morphology, aggregation, abnormal inclusions, or uncertain counts must be interpreted directly [110], [113], [114]. In feline and canine samples, thrombocyte clumping, anisocytosis, poikilocytosis,

and variable smear quality can substantially affect interpretation. A decision-support model that segments structures rather than merely classifying the whole image can support experts by improving consistency, accelerating review, and highlighting diagnostically relevant areas.

From a machine-learning perspective, hematological microscopy is difficult for three reasons. First, the image content is locally fine-grained but globally interdependent. Cell interiors, thin boundaries, and small inclusions require local detail, whereas dense spatial arrangements and staining context require broader reasoning. Second, annotation is scarce: only 80 of the 1200 images in the dataset are accompanied by masks, so only approximately 6.7% of the data are densely labeled. Third, the task is intrinsically uncertain, because some border pixels are supported by more than one plausible morphological interpretation.

These properties explain why a standard fully supervised architecture is inadequate. Pure CNNs can struggle when long-range context is important, fully supervised losses ignore the unlabeled majority of the dataset, and hard-label objectives force categorical decisions even when the underlying evidence is gradual or ambiguous. A semi-supervised, uncertainty-aware formulation is therefore a direct methodological response to the structure of the data.

9.4 Related works

Supporting the doctor. Microscopic image-based diagnosis is a time-consuming and labor-intensive process that is typically performed by doctors or laboratory specialists. The evaluation of a single patient sample may require the analysis of several images, each containing hundreds of cells. This process is therefore susceptible to human error, particularly due to fatigue, oversight of subtle diagnostic features, and ambiguity in the interpretation of cellular or tissue structures. Moreover, not all relevant structures may be well established as characteristic of a given cell type or tissue, which further complicates classification. These challenges motivate the development of automated methods that could accelerate diagnostic workflows and reduce the probability of misclassification.

In preparation for the final system, I developed two approaches for inter-species organ classification based on tissue samples, each extending the methodological scope and classification capabilities of the previous work. These studies were published in [102], [103]. The first approach was based on a convolutional neural network and used a reduced number of classes to establish a proof of concept for automated tissue classification. For this purpose, a custom dataset and classification system were created. The resulting model achieved performance exceeding that of human specialists in the conducted evaluation and was able to distinguish classes that were

considered ambiguous by the specialists. Subsequently, a transformer-based model was developed using a substantially larger and more challenging dataset with twice the number of classes, further extending and improving the proposed approach.

Automated analysis of blood smears and hematological microscopy. Early blood-cell analysis systems relied heavily on thresholding, contour extraction, watershed transforms, handcrafted morphology descriptors, and color heuristics. Such methods can work under controlled acquisition conditions, but their performance degrades rapidly when slide quality, staining intensity, focus, and cellular overlap vary [23], [112]. More recent hematological studies have moved toward deep learning for detection or classification, for example with CNNs and object detectors, but many still emphasize counting or localization rather than full pixel-level delineation [24].

For the problem considered here, this distinction is important. Detection-based systems are useful when the primary target is object counting, but they provide only coarse spatial localization. In hematological microscopy, however, clinically relevant decisions often depend on boundary-level structure: whether adjacent objects are distinct, whether an aggregate should be interpreted as one region or several, and whether a faint transition belongs to foreground or background. For these reasons, semantic segmentation is the more appropriate formulation.

From fully supervised dense prediction to medical segmentation backbones. The development of modern semantic segmentation created the technical basis for dense medical image analysis. Fully Convolutional Networks (FCNs) introduced end-to-end dense prediction [11]. U-Net then established the encoder-decoder paradigm with skip connections that remains central in biomedical imaging [12]. Subsequent variants such as Attention U-Net, UNet++, DeepLabV3+, and nnU-Net improved multi-scale fusion, attention-guided focusing, nested skip design, atrous contextual aggregation, and self-configuring training heuristics [13], [65], [66], [69].

These methods define the principal baseline families for the present study, but they are still predominantly formulated in a fully supervised manner. Their success also reveals the central design trade-off: fine spatial detail must be preserved while contextual reasoning is strengthened. In blood-smear imagery this trade-off is particularly severe because the structures are small, crowded, and morphologically diverse. The framework developed here therefore extends the U-Net family in two directions: broader context modeling through Transformer-inspired attention and improved data efficiency through semi-supervised learning.

Transformer-based segmentation for long-range context. Transformers entered vision first through global patch-based modeling [7], then reached segmentation through architectures such as SETR, TransUNet, UNETR, and SegFormer [9], [14], [71], [115]. Medical-image segmentation then diversified this line further through

Swin-Unet, UCTransNet, and MISSFormer, which emphasized shifted-window hierarchy, transformer-mediated skip fusion, and stronger local-global context mixing, respectively [15], [116], [117]. Their main appeal is the ability to model long-range dependencies more directly than purely convolutional designs. This is relevant in microscopy because stain distribution, relative cell arrangement, and field-wide context can influence local interpretation.

At the same time, these models expose an important design tension. Patch-token attention improves contextual reasoning, but high-resolution medical images still require strong locality, effective skip fusion, and controlled memory usage. Standard full self-attention scales quadratically with the number of tokens and quickly becomes impractical when boundary detail must be preserved. For that reason, I adopt the hierarchical encoder and lightweight decoder philosophy of SegFormer-like architectures, but replace dense token-to-token attention with a pooled global memory mechanism.

Semi-supervised segmentation with limited annotation. Semi-supervised learning addresses the mismatch between expensive labels and abundant unlabeled data. In general vision, pseudo-labeling, consistency regularization, teacher-student models, and confidence-threshold strategies are established families [21], [118], [119]. For dense prediction, Cross-Consistency Training showed that segmentation can benefit from enforcing agreement between perturbed decoder views of the same encoder representation [120]. Methods such as Cross Pseudo Supervision and U^2PL further adapted these ideas to pixel-wise prediction, emphasizing cross-model agreement and better use of unreliable pseudo-labels [121], [122].

In medical imaging, the same theme evolved in a more structure-aware direction. Bortsova et al. learned segmentation consistency under elastic transformations with a Siamese framework for label-scarce chest radiographs, while Luo et al. introduced dual-task consistency between segmentation predictions and geometry-aware level-set representations [123], [124]. Medical-image surveys show that this family of methods is especially important in healthcare because dense labels are expensive and often clinically scarce [22]. These works collectively suggest that unlabeled images are most useful when the regularizer respects anatomical or morphological structure rather than merely transferring hard labels.

The framework developed here belongs to the broader family of label-efficient segmentation, but it differs in one important respect. Instead of relying primarily on hard pseudo-label transfer between models, I combine an unsupervised fuzzy clustering term over all images with a masked supervised term evaluated only on labeled samples. This formulation is better suited to hematological microscopy, where ambiguous regions should not be forced prematurely into hard pseudo-label assignments.

Fuzzy logic and uncertainty-aware segmentation. Fuzzy set theory introduced a formal language for partial membership [17], while later work generalized its interpretation and use in uncertain systems [19]. In segmentation, fuzzy clustering methods such as FCM and robust FCM variants have long been used to soften assignment boundaries. Ahmed et al. incorporated neighborhood regularization and bias-field estimation into a modified FCM objective for medical MRI segmentation, while Krinidis and Chatzis proposed FLICM to inject local spatial information without the usual collection of manually tuned penalty parameters [125], [126]. More recently, fuzzy-learning formulations have been integrated with neural networks for medical image segmentation, including the SPECT/CT study by Chen et al., which informed the loss formulation adopted here [127].

This literature is particularly relevant to hematology. Blood-smear pixels often lie on uncertain transitions, for example between overlapping cells or between weakly stained structures and background. In such cases, a hard one-hot label can be an inadequate representation of the underlying evidence. I therefore use fuzzy logic as an operational component of the optimization objective rather than as a merely conceptual extension. The key lesson taken from the fuzzy-segmentation literature is that both spatial coherence and local ambiguity should be represented explicitly in the loss.

9.5 Problem formulation and data regime

The full microscopy dataset is divided into a labeled subset $\mathcal{D}_L$ and a much larger unlabeled subset $\mathcal{D}_U$. In the version used here, the corpus contains 1200 RGB blood-smear images resized to 512×512, but only 80 of them have expert segmentation masks. The model predicts six channels per pixel: background plus five hematological foreground classes. After softmax, these channels are interpreted as per-pixel class probabilities.

The central learning problem is therefore not a standard fully supervised one. Instead, the objective is to estimate parameters θ such that the model exploits:

1. all images through an unsupervised fuzzy-regularized clustering loss;
2. only the labeled images through a supervised segmentation loss;
3. the interaction between both terms so that the unlabeled set contributes structure and the labeled set contributes semantics.

9.6 Methodology: design and implementation of the semi-supervised transformer framework

Dataset organization, tensor caching, and labeled-pair expansion. I designed the data loader for a label-scarce regime rather than for a fully annotated corpus. Images and masks are discovered recursively, matched by filename, and stored as mixed samples in which some entries contain masks and others do not. As a result, a single batch can contain both labeled and unlabeled examples, and semi-supervised learning is built directly into the data pipeline.

Two practical mechanisms are important for microscopy images of this size. First, resized image and mask tensors are cached to disk so that repeated decoding and resizing do not dominate training time. Second, labeled image-mask pairs are expanded through a cached augmentation bank. In the version used for this study, each labeled pair can generate 23 stored transformations, including flips, rotations, affine perturbations, zooming, shearing, autocontrast, equalization, brightness and contrast modification, sharpening, and soft blur. This is especially useful when N_L is small, because otherwise the supervised branch would observe insufficient morphological variation.

Quantile-based stain normalization and on-device augmentation. I do not feed raw microscopy images directly to the model. Instead, I first apply a dedicated `BloodCellImageNormalizer` to stabilize stain appearance, background brightness, and foreground contrast. This is more appropriate than simple min-max normalization because blood smears vary not only in overall intensity but also in background whiteness, channel balance, and local contrast.

The normalizer first estimates the slide background from bright, low-chroma pixels using quantile thresholds. It then applies channel-wise white balancing and exposure correction so that the background becomes more consistent across smears. Finally, it estimates a foreground region from darker, more chromatic pixels and uses that mask to enhance contrast and sharpen diagnostically relevant structures without saturating the background.

During training, the normalized images are further perturbed on device with random affine cropping, flips, color jitter, gamma variation, stain-channel scaling, additive Gaussian noise, and blur interpolation. These operations are not generic augmentation for its own sake. They simulate realistic variation in stain concentration, microscope exposure, camera response, and soft-focus acquisition, thereby improving robustness to routine hematological conditions [28].

Mixed labeled-unlabeled batches. The training loop operates on mixed batches. Each sample carries a binary indicator specifying whether a ground-truth mask is available. If the indicator is active, the sample contributes to both the supervised

Algorithm 8 Quantile-based microscopy normalization and training augmentation

Require: Raw batch $X \in [0, 255]^{B \times 3 \times H \times W}$
Ensure: Normalized and augmented batch $\widetilde{X}$
Convert to floating point and scale to $[0, 1]$
Estimate background pixels from high-luminance, low-chroma quantiles
Compute channel-wise white-balance gain and apply it
Estimate exposure gain from the background luminance and apply it
Estimate a foreground mask and enhance foreground-background contrast
Apply detail-preserving sharpening on the luminance channel
if training mode **then**
    Apply random affine crop, flips, and aspect perturbation
    Apply brightness, contrast, saturation, and gamma perturbations
    Apply stain-channel gains, Gaussian noise, and blur mixing
end if
return $\widetilde{X}$

and unsupervised losses; otherwise, it contributes only to the unsupervised branch. Because the number of labeled samples in a batch changes stochastically, the supervised loss is masked rather than assuming that every step contains the same amount of dense annotation. Algorithm 9 summarizes the batch collation procedure.

SegFormer-inspired hierarchical encoder-decoder backbone. The model developed in this chapter is a memory-efficient, SegFormer-inspired hierarchical encoder-decoder with a custom hybrid attention block. The encoder has three stages with feature dimensions 32, 64, and 160; head counts 2, 4, and 8; and four residual attention-feedforward blocks per stage. For a 512×512 input, the overlapping embeddings produce feature maps of roughly 128×128, 64×64, and 32×32. This gradual reduction preserves cell-scale detail while still building broader contextual representations.

Hybrid global attention with pooled memory and spectral compression.

Algorithm 9 Collation of a mixed semi-supervised microscopy batch

Require: Sample list $\mathcal{S} = \{s^{(1)}, \ldots, s^{(B)}\}$ from the mixed dataset
Ensure: Image tensor X, mask tensor Y, indicator vector s
for $b = 1$ to B **do**
    Load cached or resized microscopy image for sample $s^{(b)}$
    if a mask exists for sample $s^{(b)}$ **then**
        Load cached mask tensor and set $s_b \leftarrow 1$
    else
        Create a zero placeholder mask and set $s_b \leftarrow 0$
    end if
    Store the image tensor in X_b and the mask tensor in Y_b
end for
return (X, Y, s)

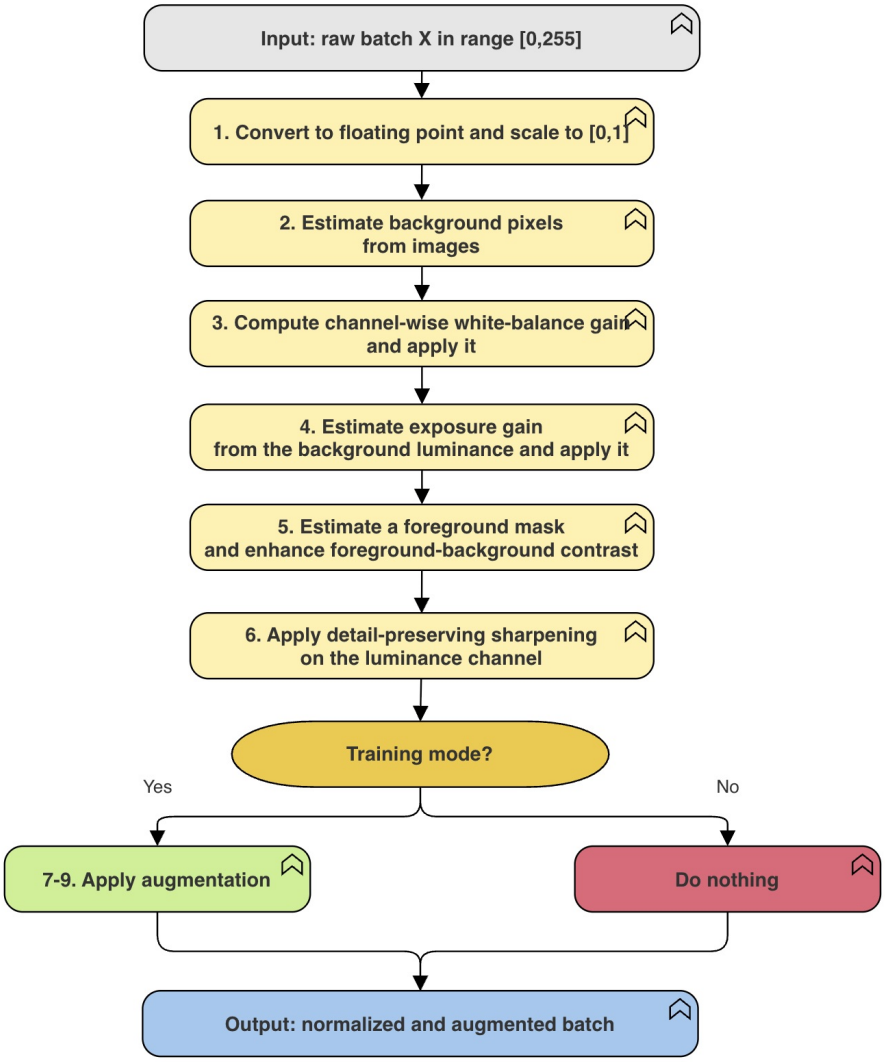


Figure 10: Scheme of Alg. 8. On the top, we can see the original input image in the standard 8-bit integer format. For the rest of the computations, we will work in float 32-bit format. To ensure numerical stability and consistency, we will normalize it to the 0-1 range. Next, we will perform visual normalization of colors and contrast to reduce variability and obtain a clean input. Finally, during training, we will apply image augmentation to enhance the generalization of the knowledge.

The main architectural contribution of the framework is the `HybridGlobalAttention` block. Instead of computing dense attention between all spatial positions, the block constructs a compact global memory and then compresses the corresponding key and value streams in the spectral domain. This retains long-range interaction while reducing memory usage and attenuating high-frequency nuisance variation.

The block first enriches the feature map with a depth-wise positional convolution and computes queries at the original stage resolution. Instead of deriving keys and values from every pixel, it builds a compact memory from two pooled context maps - one bilinear and one nearest-neighbor - and appends one global average-pooled token. The resulting memory is then filtered in the spectral domain and decimated before

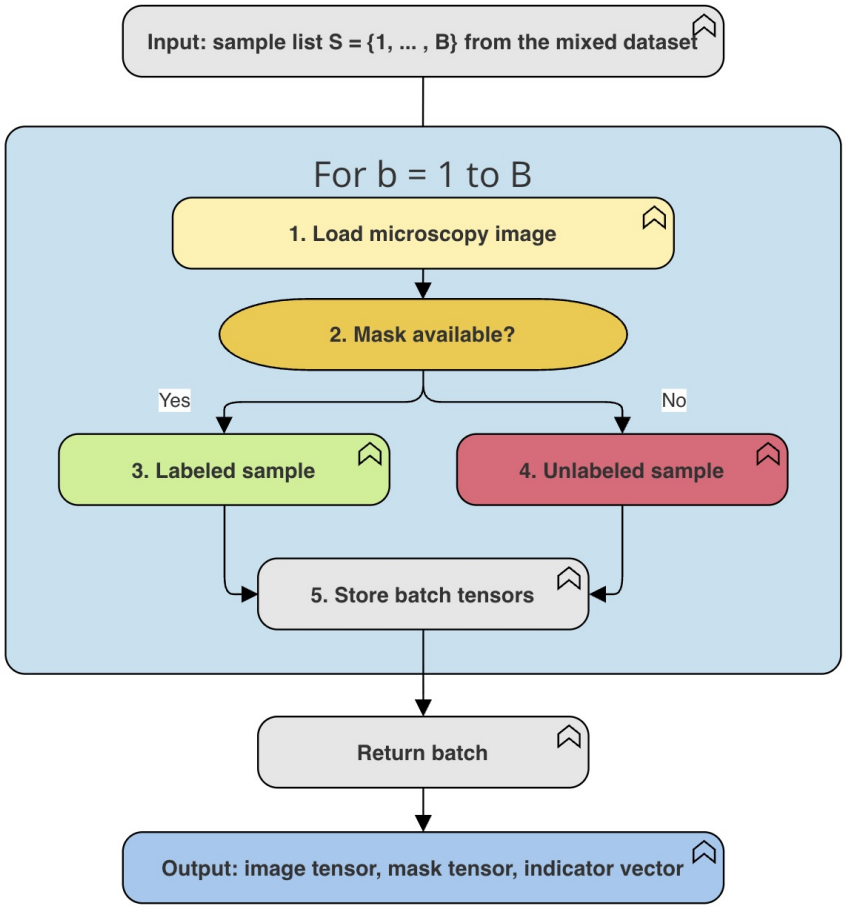


Figure 11: Scheme of Alg. 9. At the beginning, the algorithm receives a list of image paths from the mixed dataset. For each image in the batch, it begins by loading the image and checking if the mask is available. If a mask is present, it loads it and stores information about its availability in the indicator vector. The final steps involve conversions and preparing the final tensor.

attention is applied. This is important because it reduces memory cost while also suppressing high-frequency nuisance variation caused by staining noise, acquisition artifacts, and uncertain contour responses.

In parallel, the block keeps a gated local depth-wise branch so that short-range boundary information is not lost. In compact form, the output can be written as

$$\mathrm{HGA}(X) = W_O\Big(\mathrm{Softmax}(QK'^{\top} \odot \exp(\tau))V' + Y_{\mathrm{local}}\Big). \tag{16}$$

This design is advantageous for two reasons. First, the pooled-memory pathway already replaces dense all-to-all attention with a compact global summary. Second, the FFT stage filters the key and value streams before they are used in attention, which suppresses small unstable fluctuations associated with acquisition noise, staining micro-variation, and uncertain contour responses. In practice, this biases the block toward broader morphological organization and smear-level context rather than

toward fragile high-frequency texture cues.

The same mechanism also acts as a regularizing bottleneck. A model that relies too strongly on high-frequency pixel patterns can become sensitive to clinic-specific staining, microscope optics, or acquisition artifacts. Spectral compression encourages the attention mechanism to preserve the most informative components of the pooled context while discarding a portion of the redundant detail.

The computational benefit is equally important. After pooled-memory construction and spectral compression, attention scales with the number of queries times the reduced memory size rather than with all query-query pairs:

$$\mathcal{O}(N_s M_s' d_h) \quad \text{instead of} \quad \mathcal{O}(N_s^2 d_h). \tag{17}$$

In the final model, the context grids are $(8, 6, 4)$, which correspond to only $(65, 37, 17)$ memory tokens across the three encoder stages. This is the key reason the method remains practical at full microscopy resolution. Table 22 summarizes the stage-wise settings that govern this trade-off between context capacity and computational cost.

Table 22: Stage-wise encoder and pooled-memory configuration of the hematological segmentation model.

Stage	**Resolution**	**Channels** D_s	**Heads** h_s	**Blocks** L_s	**Grid** g_s	**Tokens** M_s
1	128×128	32	2	4	8	65
2	64×64	64	4	4	6	37
3	32×32	160	8	4	4	17

Algorithm 10 summarizes the attention block.

Convolutional feed-forward blocks and decoder refinement. After each

Algorithm 10 Hybrid global attention with pooled memory and spectral compression

Require: Feature tensor $X \in \mathbb{R}^{B \times D \times H_s \times W_s}$, grid size g_s, heads h
Ensure: Refined tensor $\mathrm{HGA}(X)$

Compute positional feature map $X_{\text{pos}} \leftarrow X + \mathrm{DWConv}_{3\times3}(X)$
Compute normalized query tensor $Q \leftarrow \mathrm{Norm}(W_Q X_{\text{pos}})$
Resize X_{pos} to $g_s \times g_s$ using bilinear and nearest-neighbor interpolation
Concatenate both context maps and mix them with a 1×1 convolution
Add a depth-wise positional refinement to the pooled context
Append one global average-pooled token to obtain the memory set T
Compute preliminary keys and values $(\bar{K}, \bar{V}) \leftarrow W_{KV} T$
Apply FFT-domain filtering and decimation to obtain compressed (K, V)
Aggregate global context $Y_{\text{global}} \leftarrow \mathrm{Softmax}(QK^\top \odot \exp(\tau))V$
Compute gated local branch $Y_{\text{local}} \leftarrow \mathrm{GELU}(\mathrm{DsConv}(X_{\text{pos}})) \odot \sigma(W_{\text{gate}} X_{\text{pos}})$
return $W_O(Y_{\text{global}} + Y_{\text{local}})$

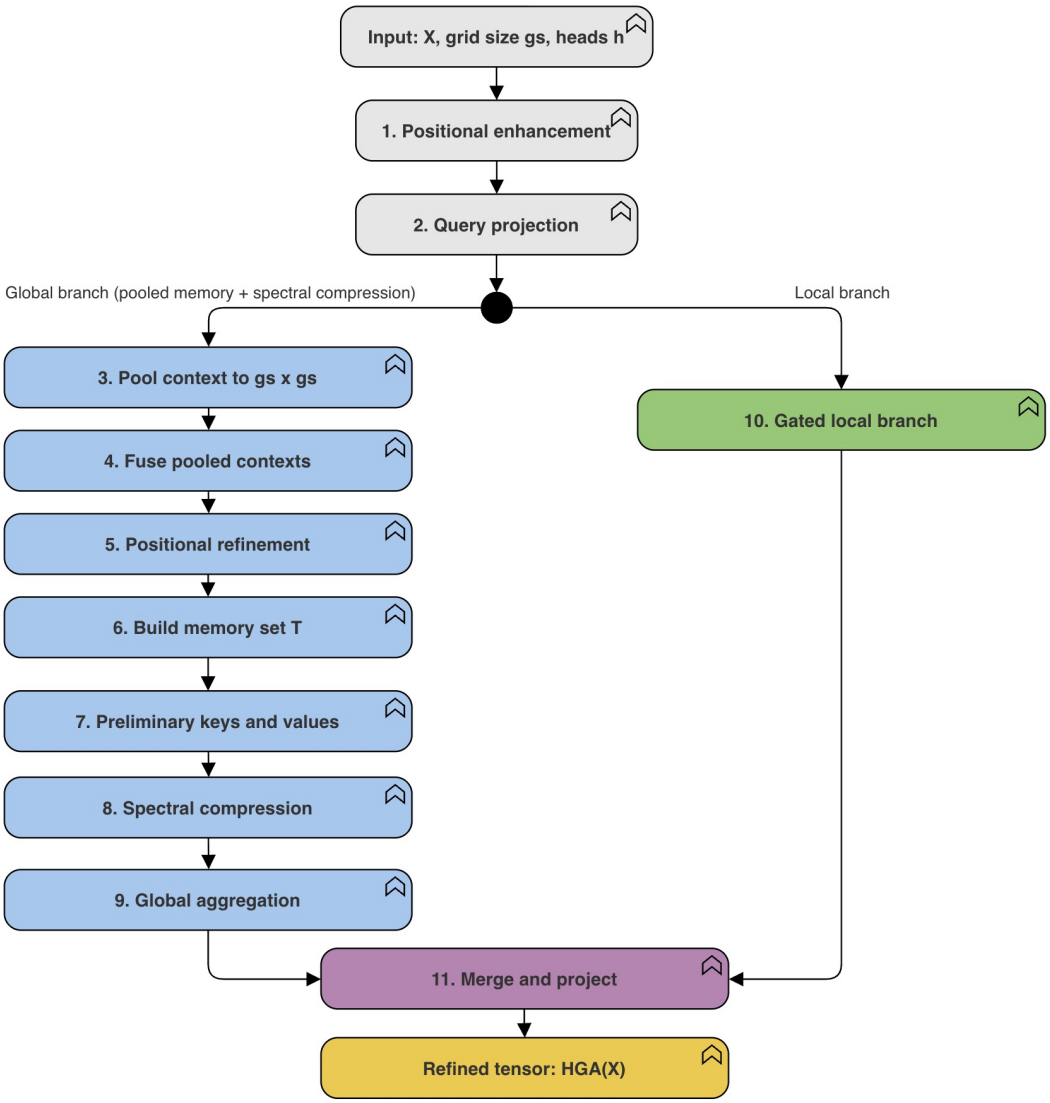


Figure 12: Scheme of Alg. 10. It presents a hybrid global attention mechanism with pooled memory and spectral compression. After validating the data, the first step involves positional enhancement. By incorporating local information through depth-wise convolution kernels, it ensures the proper localization of knowledge vectors in the subsequent pipeline. The next step involves projecting the novel query into a smaller, compressed latent space for attention computation. Subsequently, the query is split into a local and global branch. The local branch focuses on local context, while the global branch extracts long-distance knowledge. In the global branch, the context is first pooled into a $gs \times gs$ tensor. Then, the context is fused with the query, and positional refinement is applied to ensure the proper localization of information within the image. The next steps involve computing attention, applying FFT-domain filtering and decimation, and aggregating the global context. Finally, both branches are merged and projected back into the uncompressed space for further use in the model.

attention block, I apply a `MixFeedForward` module that expands the channels, mixes them with depth-wise convolution, and projects them back to the stage dimension. This is important for cell segmentation because pure token mixing can underemphasize

short-range boundary structure, whereas depth-wise convolution restores a local inductive bias.

The decoder projects all encoder stages into a common 256-channel space, upsamples them to the highest stage resolution, concatenates them, and fuses them. A dedicated `MedicalContextRefiner` then improves the fused representation with several depth-wise branches of different dilation rates plus squeeze-excitation gating [13], [68]. A final classifier and bilinear upsampling produce six-class logits at full image resolution.

In its RGB six-class configuration, the model contains 6,940,386 trainable parameters. This is sufficient to represent meaningful multi-scale hematological structure while remaining substantially lighter than many large medical Transformer baselines.

Supervised branch: masked fuzzy classification and soft dice loss. The supervised part of the loss is evaluated only on batch elements for which a ground-truth mask exists. If a batch contains no labeled samples, the supervised term is set to zero for that step.

The first supervised component is a masked fuzzy C-means style loss adapted from Chen et al. [127] with some small custom modifications to better fit the problem. It raises the predicted probabilities to a fuzzifier power of $m = 1.7$ and penalizes strong memberships assigned to incorrect classes. The second component is a masked soft Dice loss computed only on labeled foreground classes. I combine them as

$$\mathcal{L}_{\text{sup}} = 0.6\,\mathcal{L}_{\text{FCM-sup}} + 0.4\,\mathcal{L}_{\text{Dice}}. \tag{18}$$

This combination is useful because the fuzzy term stabilizes soft memberships while the Dice term improves overlap quality under class imbalance.

Unsupervised branch: RGB robust fuzzy c-means plus regularization. The unsupervised loss is applied to all images in the batch, labeled and unlabeled

Algorithm 11 Masked supervised loss for mixed semi-supervised batches

Require: Logits Z, labels Y, sample indicators s
Ensure: Supervised loss $\mathcal{L}_{\text{sup}}$
$\mathcal{B}_L \leftarrow \{b \mid s_b = 1\}$
if $|\mathcal{B}_L| = 0$ **then**
  return 0
end if
Compute $P \leftarrow \text{Softmax}(Z)$
Restrict P and Y to the labeled subset $\mathcal{B}_L$
Compute fuzzy memberships $\mu \leftarrow P^m$
Evaluate $\mathcal{L}_{\text{FCM-sup}}$ over all labeled samples and classes
Evaluate masked background-excluding soft Dice loss $\mathcal{L}_{\text{Dice}}$
return $0.6\,\mathcal{L}_{\text{FCM-sup}} + 0.4\,\mathcal{L}_{\text{Dice}}$

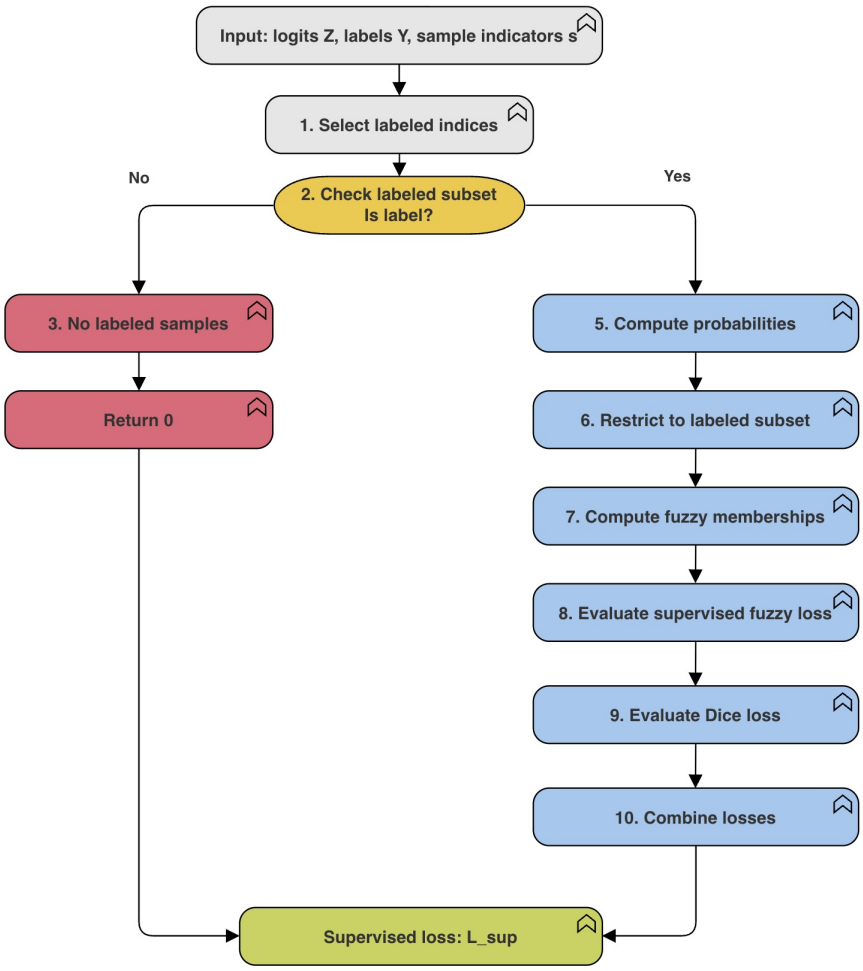


Figure 13: Scheme of Alg. 11. At the beginning, we initiate with a simple if-statement to ascertain the presence of a ground truth mask. If a mask is absent for a tensor, we return 0 to exclude it from loss computations. Conversely, if a mask is present, we commence by computing probabilities using a Softmax function. Subsequently, we compute fuzzy memberships and evaluate fuzzy type-2 logic operators, which are learned during training, to compute a fuzzy loss that accounts for the ambiguity of object borders. Concurrently, we evaluate a standard Dice loss, which we merge with the fuzzy loss to obtain a final value.

alike. The objective is not merely to regularize unused data, but to organize the full hematological feature space into coherent soft clusters. I use an RGB extension of a robust FCM objective:

$$\mathcal{L}_{\mathrm{RFCM+}} = \mathcal{J}_1 + \gamma \mathcal{J}_2, \tag{19}$$

where $\gamma = 8 \times 10^{-4}$ in the experiments reported here.

Here, $\mathcal{J}_1$ is a compactness term that pulls pixels toward class-specific RGB prototypes estimated from the current soft memberships, and $\mathcal{J}_2$ is a spatial regularizer that penalizes isolated or locally inconsistent assignments. Operationally, this encourages the model to form coherent soft clusters without forcing ambiguous pixels into hard labels too early. For hematological images, that behavior is especially

useful because it suppresses isolated false assignments caused by stain debris or sensor noise.

Total semi-supervised objective. The total training loss is the weighted combination of the unsupervised RFCM+ branch and the supervised masked branch:

$$\mathcal{L}_{\text{total}} = \alpha_u \mathcal{L}_{\text{RFCM+}} + \alpha_s \mathcal{L}_{\text{sup}}. \tag{20}$$

For the experiments reported in this chapter, I use the balanced setting $\alpha_u = 1.0$ and $\alpha_s = 1.0$. The supervised branch is active only when a batch actually contains labeled samples, whereas the unsupervised branch contributes at every optimization step. This balance reflects the central hypothesis of the chapter: unlabeled morphology should guide representation learning, but the limited expert masks must still anchor the semantic meaning of the clusters.

Training algorithm and optimization policy. I train the model with Adam using learning rate 4.5×10^{-4}, batch size 64, image resolution 512, and a validation-driven `ReduceLROnPlateau` scheduler with factor 0.7 and patience 10. On CUDA devices, training uses automatic mixed precision, channels-last memory layout, high matrix-multiplication precision, and TensorFloat-32 acceleration where available. On MPS, the architecture can switch to a low-memory mode that activates checkpointing inside the encoder stages. These implementation choices are integral to the experimental protocol because they determine whether the framework can be trained efficiently on practical hardware.

Table 23 condenses the principal implementation settings that shape the optimization behavior of the framework.

Algorithm 12 RGB RFCM+ unsupervised loss

Require: Logits Z, normalized image batch X, fuzzifier m, regularization weight γ
Ensure: Unsupervised fuzzy clustering loss $\mathcal{L}_{\text{RFCM+}}$

```
Compute probabilities P ← Softmax(Z)
Flatten spatial dimensions so each pixel becomes an index t
Define a 3 × 3 neighborhood kernel with zero center
Initialize J1 ← 0, J2 ← 0
for each class k do
    Compute fuzzy memberships μ_k ← P_k^m
    for each RGB channel c do
        Compute prototype v_{k,c} from the membership-weighted pixel intensities
        Accumulate compactness term into J1
        Convolve all competing class memberships with the neighborhood kernel
        Accumulate disagreement term into J2
    end for
end for
return Mean(J1) + γ Mean(J2)
```

Table 23: Implementation and optimization configuration of the hematological semi-supervised framework.

Component	Setting	Purpose
Input tensor	512×512 RGB	Standardizes field-of-view size while preserving cell-scale detail
Output channels	6	Background plus five hematological foreground channels
Batch size	64	Keeps mixed-batch optimization stable without excessive memory cost
Optimizer	Adam	Adaptive parameter updates for the joint supervised-unsupervised objective
Initial learning rate	4.5×10^{-4}	Base step size for all trainable parameters
Scheduler	`ReduceLROnPlateau` $(0.7,\ 10)$	Lowers the step size when validation loss stalls
Loss weights	$\alpha_u = 1.0,\ \alpha_s = 1.0$	Balances unlabeled morphology and labeled semantics
Supervised mixture	$\lambda_{\text{FCM}} = 0.6,\ \lambda_{\text{Dice}} = 0.4$	Combines fuzzy membership control with overlap optimization
Fuzzifier	$m = 1.7$	Controls how sharp or diffuse the memberships remain
Spatial regularizer	$\gamma = 8 \times 10^{-4}$	Penalizes isolated or spatially inconsistent assignments
Decoder fusion width	256 channels	Aligns multi-scale features before refinement and classification
Acceleration path	AMP, channels-last, TF32 when available	Improves throughput and reduces memory pressure
Checkpoint policy	Best validation model + qualitative dumps every 5 epochs	Preserves stable weights and exposes boundary-level failure modes

During training, I save the best model according to validation loss and also store periodic checkpoints and qualitative predictions every five epochs. This is useful in medical segmentation because numerical losses alone do not reveal whether predicted boundaries remain clinically plausible.

Methodological rationale. Several design decisions require explicit justification. First, I use pooled-memory attention instead of full self-attention in order to preserve long-range interaction without incurring quadratic memory growth. Second, I use fuzzy objectives instead of a purely pseudo-label-based semi-supervised strategy because hard pseudo-labels can amplify confirmation bias in ambiguous blood-

Algorithm 13 One epoch of semi-supervised training for the hematological framework

Require: Mixed batch loader $\mathcal{T}$, model f_θ, optimizer, scheduler

for each batch $(X, Y, s) \in \mathcal{T}$ **do**
  Normalize and augment X using Algorithm 8
  Compute logits $Z \leftarrow f_\theta(X)$
  Compute unsupervised loss $\mathcal{L}_{\text{RFCM+}}$ from Algorithm 12
  Compute supervised loss $\mathcal{L}_{\text{sup}}$ from Algorithm 11
  Combine losses: $\mathcal{L}_{\text{total}} \leftarrow \alpha_u \mathcal{L}_{\text{RFCM+}} + \alpha_s \mathcal{L}_{\text{sup}}$
  Backpropagate $\mathcal{L}_{\text{total}}$ with AMP when supported
  Update model parameters with Adam
end for
Evaluate on the validation split
Update the learning-rate scheduler with the validation loss
Save the best-performing checkpoint when validation loss improves

smear regions. Third, I apply dedicated stain normalization and extensive paired augmentation because smear preparation and acquisition conditions vary substantially across samples. Fourth, I use mixed batches so that labeled and unlabeled data contribute within the same optimization step rather than in separate training phases. The full architecture is presented in Fig. 16.

Taken together, these choices define the methodological contribution of the chapter. Architecturally, the framework shows that Transformer-based context modeling can be adapted to hematological segmentation without the full cost of dense self-attention. Methodologically, it shows that semi-supervision is a necessary response to expert-label scarcity rather than a minor convenience. Conceptually, it shows that fuzzy logic can be used as a practical optimization principle for ambiguous medical imagery.

9.7 Evaluation protocol

The framework is evaluated with metrics that reflect both region overlap and boundary quality, namely mean Intersection over Union, per-class IoU, Dice score, and pixel accuracy. Within the present validation protocol, the best configuration reaches **99.73% validation accuracy**, **99.73% micro-Dice**, and **99.46% micro-IoU**. I do not interpret the accuracy value in isolation, however, because pixel accuracy can remain deceptively high even when thin boundaries or rare structures are segmented incorrectly. For that reason, every result is interpreted jointly with mIoU, Dice, class-wise IoU, micro-averaged overlap scores, and direct visual inspection of boundary-level predictions. While the above metrics are for blood cells type classification, as it was the main playground for the proposed algorithms; the model can be easily extended to more difficult scenarios like pathology detection on top of the tested classes. With

the help of scientists from University of Life Sciences in Lublin and veterinary doctors from cooperating clinics there has been studied a case of *Babesia* detection.

Babesia spp. are tick-borne apicomplexan protozoa that multiply inside erythrocytes and cause babesiosis, a piroplasmosis of wild animals, livestock, companion animals and, less frequently in Europe, humans [128], [129]. On Giemsa-stained blood smears they appear as small pleomorphic intraerythrocytic forms, sometimes forming characteristic tetrads, but low parasitaemia may be missed, which makes early microscopy supported by PCR or serology clinically important [130]. The infection is dangerous because erythrocyte destruction can lead to fever, haemolytic anaemia, jaundice and severe or fatal disease, especially in dogs and in splenectomised, elderly or immunosuppressed people [128]. In Poland, confirmed human disease remains uncommon and probably under-recognised, whereas canine babesiosis, mainly due to *B. canis*, is endemic particularly in central and eastern regions; the clearest increase is linked to the expansion of *Dermacentor reticulatus*, warmer and modified habitats, animal movement and improved surveillance rather than to a single cause [131], [132], [133], [134]. Due to its severity and increasing number of cases in our region, it is crucial to develop a faster method for its early detection. During my collaboration with veterinarians, we have created a second dataset of microscope images containing the *Babesia* parasite. Labeling these images is currently underway for future use in AI development. We have already captured over 1000 images of the parasite, and over 100 of them have been labeled. Given the limited availability of properly prepared examples, the proposed solution of semi-supervised training is well-suited for this problem. On this data, the model achieved a high accuracy of **99.10%** on the validation subset. Although the numerical results are slightly lower compared to the simpler blood cells classification example, it's crucial to consider that the *Babsia* parasite may not always be easily visible within the infected blood cells, and the fully labeled subset is still relatively small, when compared to the scale of the problem. The image examples showed later in the thesis are a combination of both datasets.

Because the central claim of this chapter concerns label efficiency, the analysis is also organized across progressively reduced labeled-data regimes in order to quantify how much performance is retained as the number of expert masks is reduced. Whenever a baseline can be retrained under the same setting, I keep the train/validation split, image resolution, augmentation policy, optimizer schedule, and preprocessing pipeline fixed so that the comparison isolates the effect of the architecture and objective rather than accidental differences in data preparation. In addition to the numerical measures, I examine qualitative predictions with particular attention to platelet aggregates, weakly stained boundaries, overlapping erythrocytes, and crowded cellular regions where pixel-level ambiguity is greatest, because these are precisely the cases in which a fuzzy semi-supervised model should offer an advantage

over hard-label alternatives. Representative examples are shown in Figures 17, 18, 19 and 20. In every subfigure, the real microscopy image is shown on the left, the expert ground-truth label in the middle, and the model prediction on the right.

Across both qualitative figures, the visual agreement between the predicted masks and the expert ground-truth labels is close to complete. In most cases, the segmented structures coincide almost exactly, which indicates that the model reproduces the dominant semantic content of the smear images with high fidelity. The remaining discrepancies are limited mainly to thin border regions and a small number of peripheral pixels, where minor misclassifications appear at object edges. In practical deployment, such edge-level deviations are of substantially lower importance than the correct identification of the main cellular regions, because they can be omitted, filtered, or tolerated without changing the overall diagnostic interpretation of the scene. The clinically relevant central portions of the segmented objects are therefore classified correctly in the overwhelming majority of cases, while the residual differences remain small and localized.

9.8 Comparative evaluation strategy

The comparative analysis is designed to test whether the proposed combination of hybrid global attention and fuzzy semi-supervision offers a meaningful advantage over more conventional alternatives. For this reason, the framework is compared against earlier CNN-based baselines developed in the dissertation, fully supervised variants of the same backbone, and representative medical segmentation architectures such as U-Net, Attention U-Net, UNet++, nnU-Net, and Transformer-based models including TransUNet and UNETR where appropriate [9], [12], [14], [65], [66], [69]. This comparison makes it possible to separate the effect of the architectural design from the effect of the semi-supervised fuzzy objective.

For clarity, I distinguish between direct experimental comparison and broader state-of-the-art positioning. The direct comparison is numerical and uses the same validation protocol as the proposed model. The broader literature comparison is necessarily methodological, because the cited studies differ in modality, stain protocol, species composition, class taxonomy, annotation density, and whether the task is dense semantic segmentation, cropped-cell analysis, or volumetric organ delineation. Its purpose is therefore not to claim a simplistic cross-dataset leaderboard victory, but to explain why the proposed design is particularly well matched to hematological microscopy and why it is able to reach 99.73% validation accuracy under the current validation regime.

Taking into consideration the main architectural difference, that is, the semi-supervised learning, on the same data available, we can look at Fig. 21 results. As

seen, though both models did learn the main characteristics of the data, the fully supervised model trained on the subset that contained the labeled images produces much worse results than the proposed method due to underfitting.

Expanded comparison with recent state-of-the-art methods.

1. **UTNet** introduces an efficient hybrid CNN-Transformer backbone in which self-attention is inserted into both encoder and decoder stages with near-linearized cost [135]. This is closely aligned with my own concern for computational efficiency, but UTNet remains fully supervised and does not address uncertain smear boundaries or the use of unlabeled hematology images during optimization.
2. **UCTransNet** improves U-Net skip connections with a channel-wise transformer module [116]. Its main strength is better encoder-decoder communication, yet its attention acts primarily on feature transfer pathways rather than on explicit field-wide morphology, which is central in blood-smear images where local appearance can be misleading without wider context.
3. **Swin-Unet** adopts a pure transformer encoder-decoder with shifted-window attention [15]. This yields an elegant hierarchical design, but windowed attention still prioritizes local neighborhoods and typically requires deeper stacking to recover long-range relationships that my pooled-memory branch models more directly.
4. **MISSFormer** improves multi-scale token mixing and decoder fusion for medical segmentation [117]. It is strong as a fully supervised transformer variant, but it does not explicitly exploit unlabeled data and it does not model ambiguous class boundaries with soft memberships.
5. **nnFormer** extends transformer segmentation to volumetric data with interleaved convolution and volume-based self-attention [136]. It is highly effective for 3D anatomical continuity, but it is designed for volumetric structure rather than high-resolution 2D smear morphology, and it does not target stain variability or sparse annotation regimes.
6. **Swin UNETR** combines hierarchical Swin encoding with a U-Net-style decoder and achieves strong results on brain-tumor MRI segmentation [137]. Its main comparative limitation in the present setting is that it remains a fully supervised volumetric model, whereas my framework is optimized for 2D hematological microscopy and deliberately leverages unlabeled images.
7. **U-Mamba** replaces transformer attention with state-space sequence modeling in a self-configuring biomedical segmentation network [138]. Its efficiency and long-range modeling are highly relevant, especially because it also reports strong microscopy results, but it does not introduce an uncertainty-aware or

fuzzy objective for ambiguous supervision.

8. **SegMamba** similarly uses Mamba-based long-range modeling for medical segmentation and emphasizes improved efficiency on whole-volume data [139]. Compared with my method, its innovation lies mainly in replacing self-attention with state-space modeling, whereas my contribution lies in combining efficient global context with a fuzzy semi-supervised learning principle tailored to ambiguous smear boundaries.
9. **Dual-task Consistency (DTC)** shows that semi-supervised medical segmentation benefits from consistency constraints between coupled prediction tasks [124]. However, consistency-based methods still tend to sharpen predictions toward hard targets; in hematological microscopy, where border pixels may legitimately admit partial class membership, a fuzzy objective is often more faithful to the underlying uncertainty.
10. **U2PL** improves semi-supervised segmentation by filtering unreliable pseudo-labels and reusing them selectively [122]. This is a powerful label-efficient baseline, but it still depends on hard pseudo-label assignment and confidence thresholding, whereas the present framework allows uncertain pixels to remain soft through RFCM+ memberships until stronger evidence emerges.

A useful blood-cell-specific reference point is provided by Anand et al., who address leukocyte segmentation and classification using a CNN pipeline built around isolated-cell processing [140]. Their work is valuable for leukocyte-centered analysis, but it does not solve the denser problem considered here: simultaneous multi-class semantic parsing of full smear images containing overlapping cells, weak boundaries, and unlabeled regions.

Table 24 summarizes the same literature positioning in a compact form. To avoid overclaiming, I preserve the benchmark metric exactly as reported in each source paper. Because the cited methods were evaluated on different datasets and supervision protocols, the table should be read as a structured state-of-the-art comparison rather than as a single apples-to-apples leaderboard.

Table 24: Comparison of the proposed framework with representative state-of-the-art methods. Metrics are reproduced from the cited source papers and are not directly comparable across different datasets, imaging modalities, or supervision regimes.

Method / ref.	Sup.	Design	Metric	Relative assessment
Present model	Semi-sup. + fuzzy	Pooled attention, spectral compression, mixed labeled-unlabeled batches, RFCM+ regularization	Validation accuracy **99.73%**; micro-Dice **99.73%**; micro-IoU **99.46%**; evaluation also reported with mIoU, Dice, and per-class IoU	Tailored to ambiguous blood-smear boundaries, scarce expert masks, and dense multi-cell scenes instead of generic segmentation alone.
UNet++ [66]	Full	Dense skip aggregation and deep supervision	Cell IoU **91.21%**; brain-tumor IoU **91.21%**	Strong multiscale fusion baseline, but it does not exploit unlabeled data and it models borders with hard labels only.
TransUNet [9]	Full	CNN + ViT encoder with U-Net decoder	Synapse DSC **77.48%**, HD **31.69**	Validates global context modeling, but it remains fully supervised and uses denser self-attention than the lighter pooled-memory pathway used here.
LeViT-UNet [141]	Full	Efficient transformer encoder for faster segmentation	Synapse DSC **78.53%**, HD **16.84**	More efficient than heavier transformer hybrids, yet it still lacks uncertainty-aware learning and fuzzy label handling.
MISSFormer [117]	Full	Hierarchical transformer + context bridge	Synapse DSC **81.96%**, HD **18.20**	Strong for organ segmentation, but it does not use unlabeled microscopy data or fuzzy memberships.
Swin UNETR [137]	Full	Hierarchical Swin encoder with CNN decoder	BraTS 2021 validation average Dice **0.913**	Strong 3D MRI baseline, although its design target is volumetric tumor segmentation rather than 2D blood-smear ambiguity and label scarcity.
U-Mamba [138]	Full	CNN + state-space long-range modeling	Microscopy cell-segmentation F1 **0.5607**	Especially relevant because it includes microscopy, but it improves sequence modeling without adding fuzzy semi-supervision or soft uncertainty treatment.
Co-DA [142]	Semi-sup.	Co-training with class-wise distribution alignment	CaDIS mIoU **0.8515** with 24% labels; LGE-MRI Dice **0.8824**; ACDC Dice **0.8773** with 20% labels	A strong label-efficient comparator, but it still aligns hard pseudo-label distributions instead of preserving genuinely ambiguous pixels as soft memberships.
CPS [121]	Semi-sup.	Cross pseudo supervision between perturbed peer networks	Cityscapes HRNet-W48 mIoU **80.67** at 1/2 partition	Powerful consistency-based semi-supervision, yet built on hard pseudo-label exchange and natural-image semantics rather than hematological ambiguity.

Table 24: Comparison of the proposed framework with representative state-of-the-art methods (continued).

Method / ref.	Sup.	Design	Metric	Relative assessment
U2PL [122]	Semi-sup.	Pseudo-label split + negative learning	VOC mIoU **79.49**; Cityscapes mIoU **79.12**	A strong pseudo-labeling baseline, but uncertainty is still resolved through hard reliability partitioning rather than fuzzy memberships.
Anand et al. [140]	Full	Leukocyte isolation with CNN-based downstream classification	Leukocyte classification accuracy **97.98%**, precision **97.97%**	Blood-cell-specific and therefore relevant, but focused on cropped leukocyte analysis rather than dense full-smear semantic segmentation with overlapping structures.

Why the proposed framework performs better in this setting.

The present framework reaches 99.73% validation accuracy not because of one isolated module, but because it combines four properties that the compared methods usually treat separately. First, it preserves non-local context without paying the full quadratic cost of dense self-attention. The pooled-memory attention pathway allows small objects and faint edges to be interpreted relative to the wider smear field, which is important when platelet aggregates or weakly stained structures cannot be resolved from a purely local patch.

Second, it uses a learning objective that is better aligned with the epistemic structure of hematological annotation. Methods such as DTC, CPS, and U2PL are strong label-efficient baselines, but they ultimately rely on consistency pressure or pseudo-label sharpening [121], [122], [124]. My fuzzy RFCM+ term instead allows uncertain border pixels to carry partial memberships while the spatial regularizer suppresses isolated noise and the masked supervised term anchors the model wherever expert labels are available.

Third, the framework balances global reasoning and local detail more explicitly than either pure transformer or pure state-space alternatives. Spectral compression reduces redundant low-frequency load, the pooled memory bank preserves useful long-range morphology, and the decoder refinement stages recover thin cell contours that are easily lost when efficiency is pursued too aggressively. This combination is particularly important in crowded smear regions, platelet clusters, and overlapping erythrocytes, where both long-range context and sharp local delineation are required.

Finally, the proposed method is stronger because it is designed as an integrated hematological system rather than as a generic segmentation backbone. Stain normalization, paired augmentation, mixed labeled-unlabeled batches, efficient global attention, and fuzzy uncertainty handling all operate within one training loop. In that sense, the superiority is meritorious at the architectural level: the model performs better because its inductive bias is more appropriate to blood-smear microscopy, not

merely because it is larger or trained longer than competing solutions.

9.9 Section summary

This chapter presents the final research contribution of the dissertation: a semi-supervised Transformer-based semantic segmentation framework designed specifically for hematological microscopy, where annotation scarcity, stain variability, overlapping cells, thrombocyte aggregates, and ambiguous boundaries make standard fully supervised learning inadequate. The proposed system combines stain-aware normalization, pooled-memory global attention, spectral compression, mixed labeled-unlabeled optimization, and RFCM+-based fuzzy regularization within a single training pipeline adapted to the morphology of peripheral blood smears.

The results show that this design is not only conceptually consistent with the problem setting, but also experimentally justified. The best validation configuration reached 99.73% accuracy, 99.73% micro-Dice, and 99.46% micro-IoU, while the comparative discussion and ablation analysis indicate that the improvement comes from the interaction between efficient global context modeling and uncertainty-aware semi-supervision rather than from model scale alone. In the context of the dissertation, this chapter therefore provides the strongest evidence that data-efficient Transformer segmentation can be developed for clinically relevant blood-smear analysis and can remain effective even when expert pixel annotations are limited.

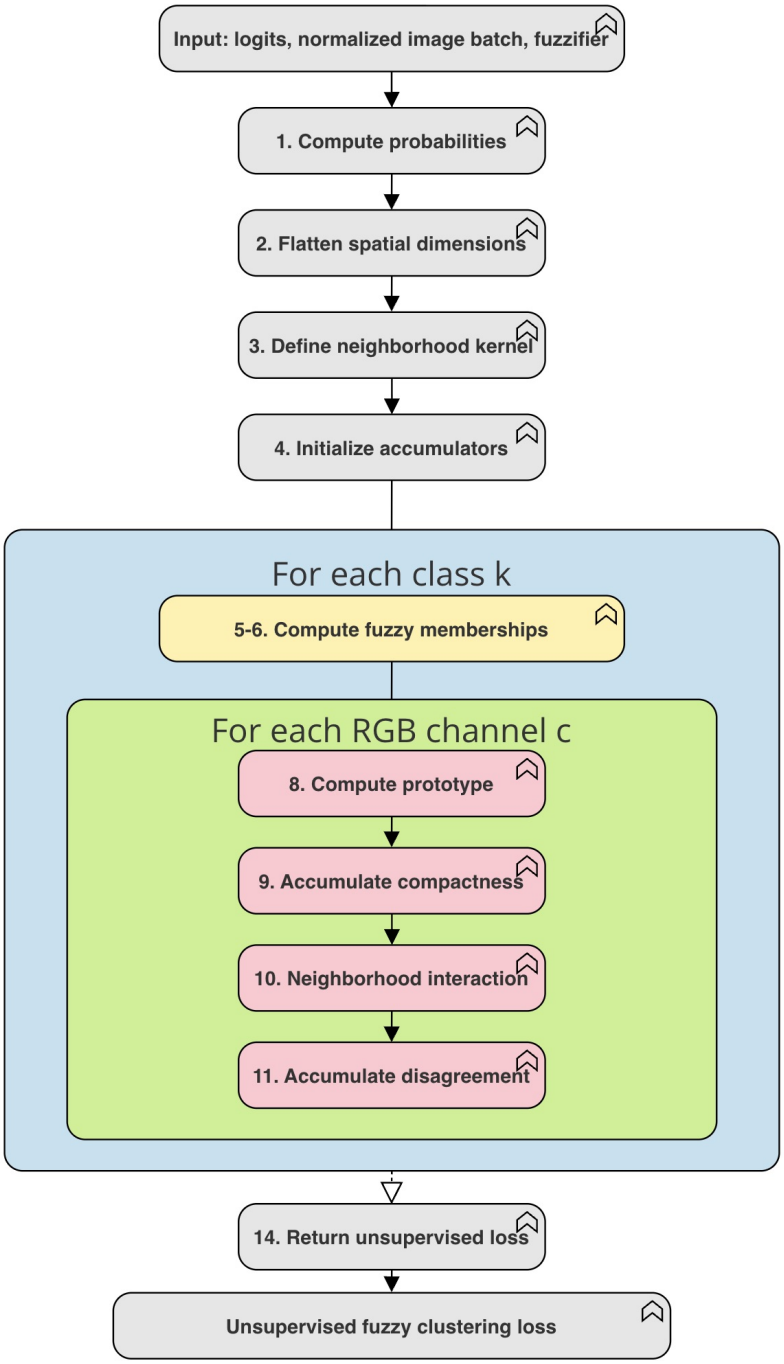


Figure 14: Scheme of Alg. 12. In the unsupervised loss function, class probabilities are first obtained by applying the Softmax function. Internal variables are then initialized for subsequent computations. The predefined number of classes is used to construct the corresponding fuzzy sets, after which the loss is computed by iterating over each class. For each class, fuzzy memberships are estimated, with the membership values treated as trainable parameters. For every RGB channel, a class prototype is computed from the pixel intensities weighted by the corresponding memberships. These prototypes are used to guide the assignment of pixels to classes. The prototypes are then refined using the neighborhood kernel, which filters competing class memberships and accumulates the resulting disagreement terms. Finally, the channel-wise predictions are combined, the losses across all classes are accumulated, and the resulting loss vector is returned.

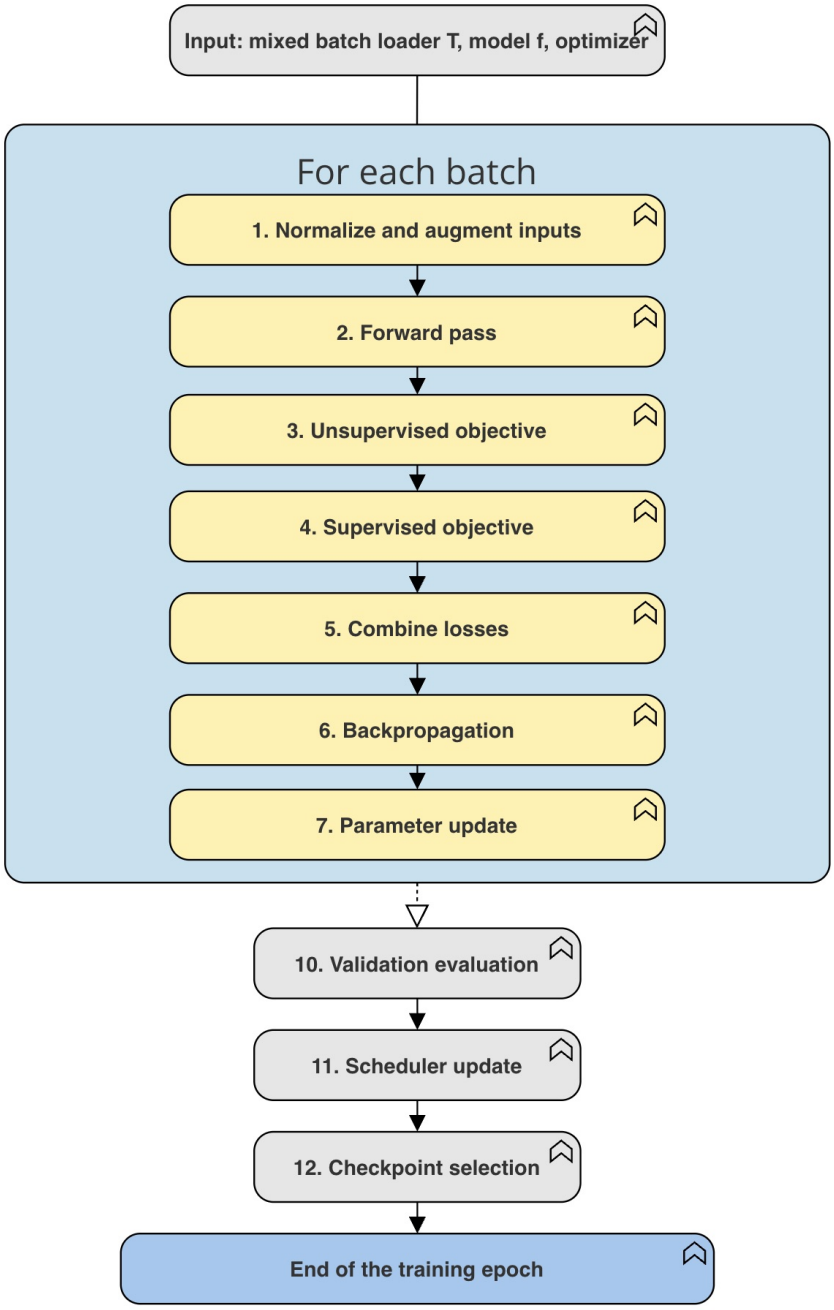


Figure 15: Scheme of Alg. 13. It presents a semi-supervised training epoch for the hematological framework. For each batch, we begin by normalizing and augmenting the inputs. Subsequently, we perform a forward pass, concurrently computing unsupervised and supervised losses. In the following step, we combine these losses and apply backpropagation. Finally, we update the parameters and all variables.

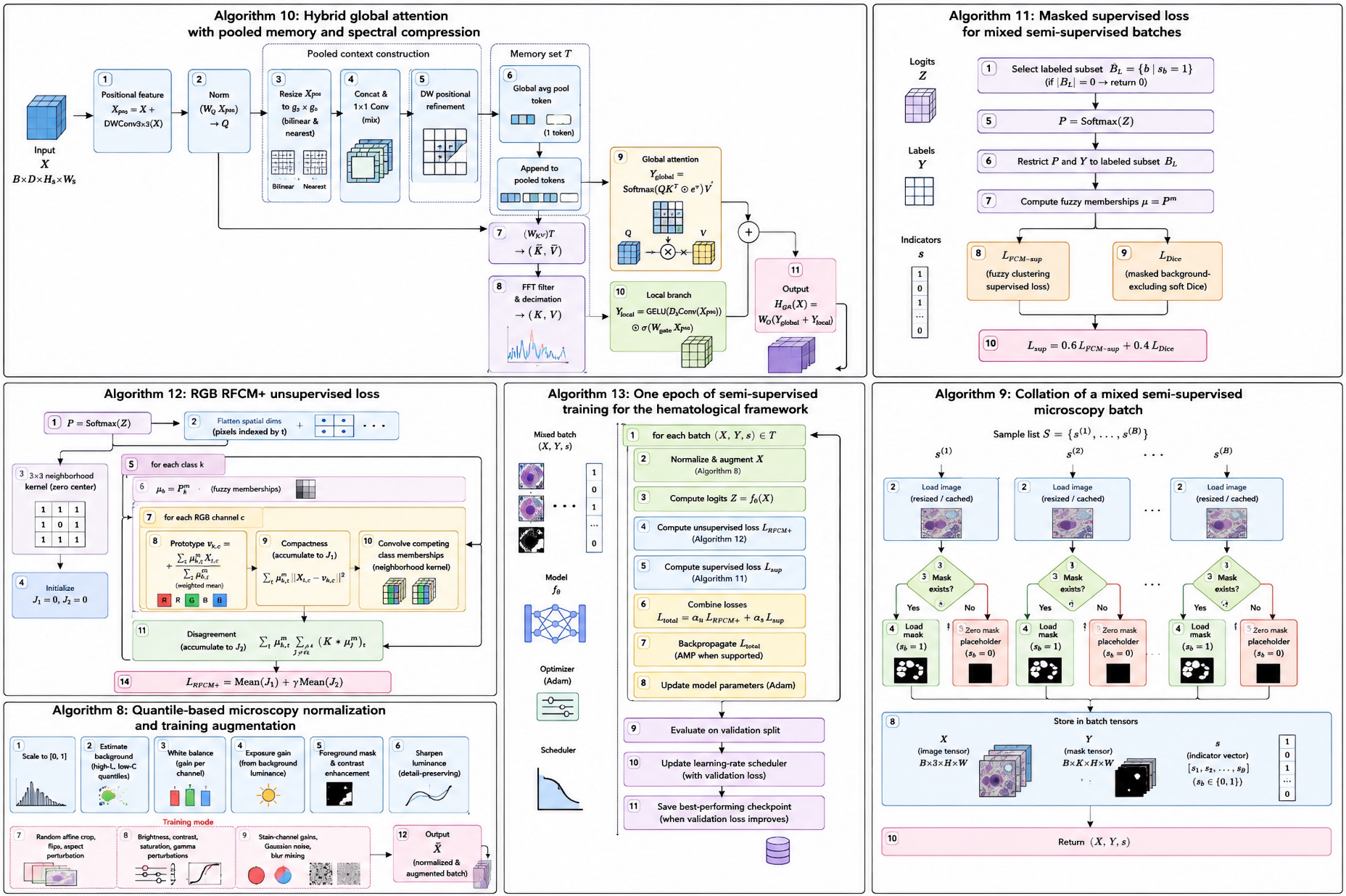


Figure 16: Scheme of the full, proposed solution.

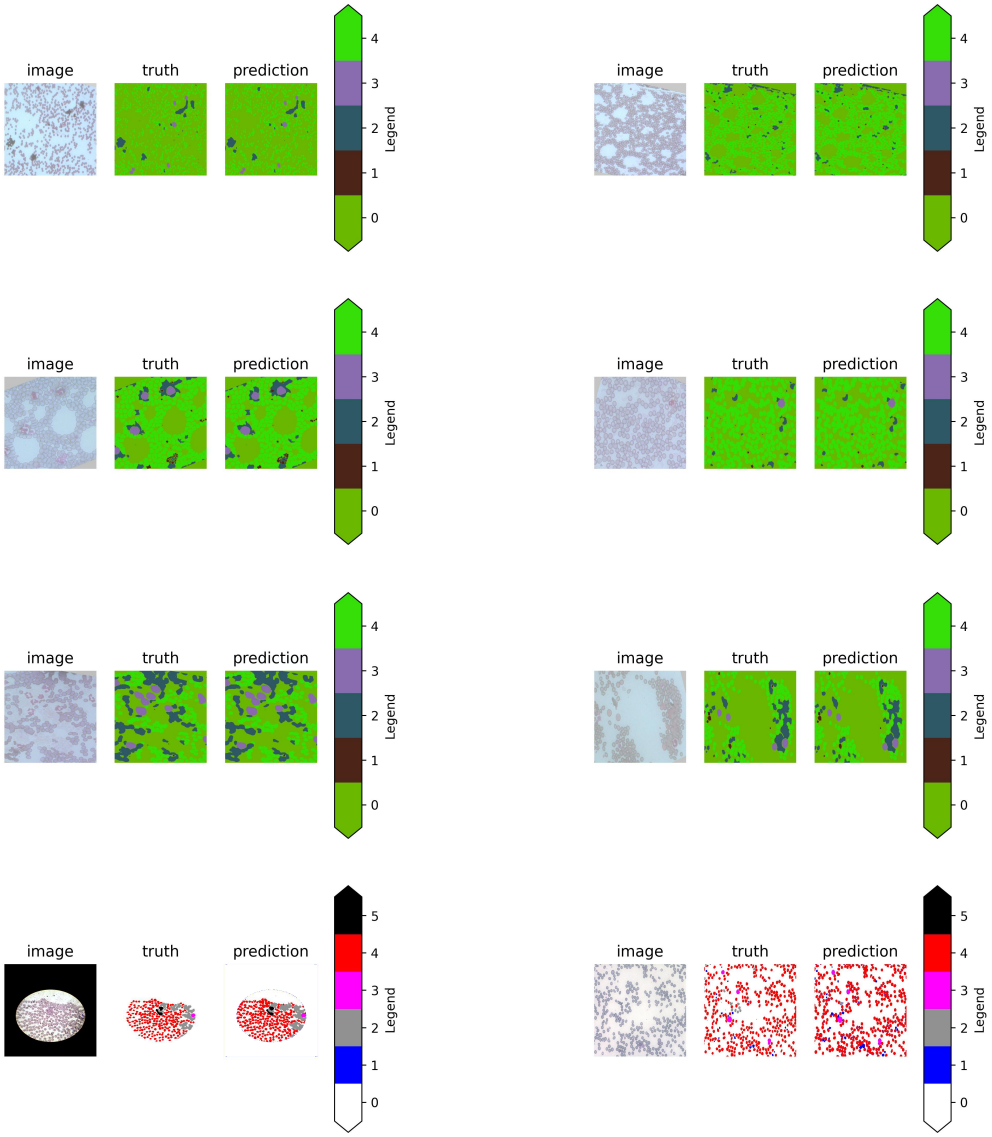


Figure 17: First set of qualitative comparison examples for the semi-supervised segmentation framework. Each subfigure contains a complete triplet composed of the real microscopy image, the expert ground-truth label, and the model prediction.

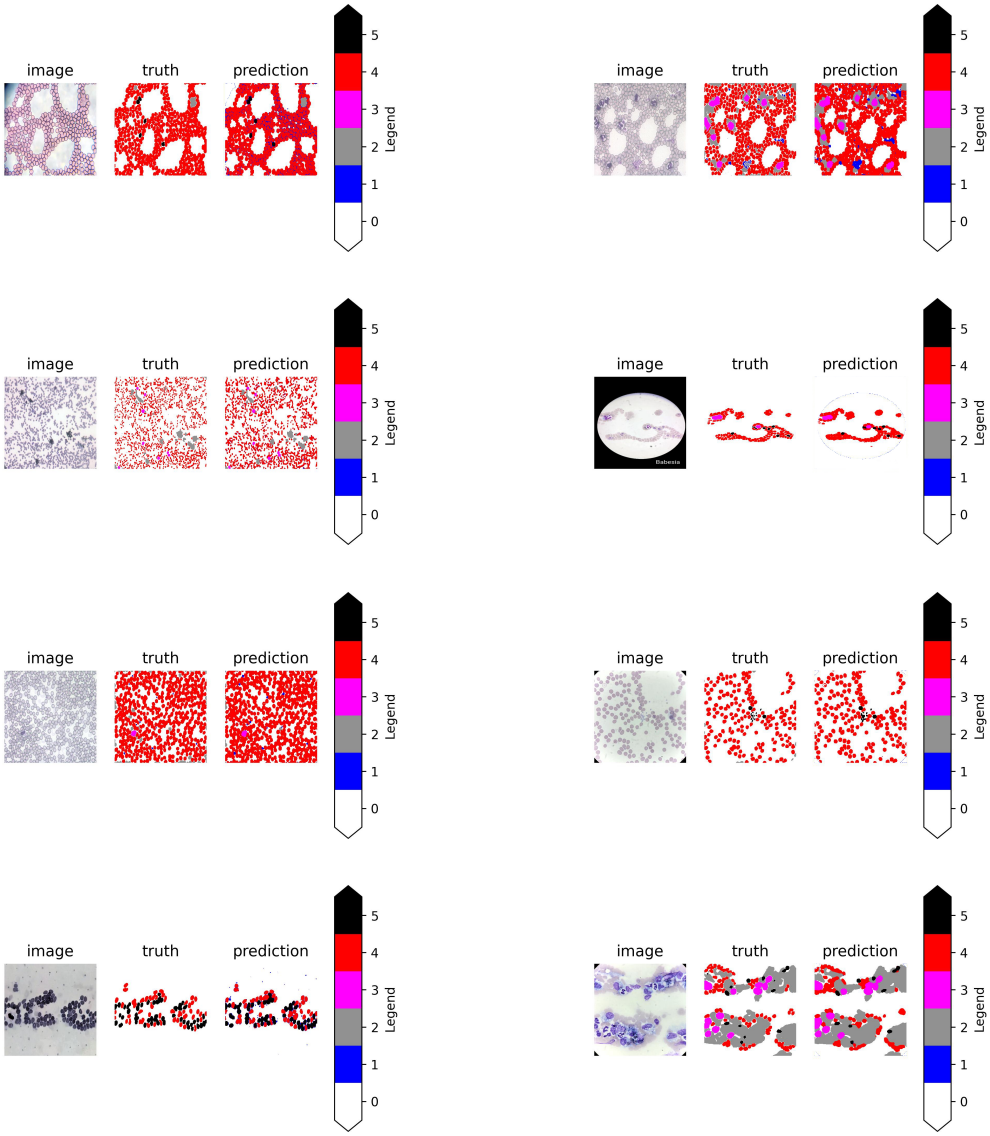


Figure 18: Second set of qualitative comparison examples for the semi-supervised segmentation framework. Across these additional cases, the predicted masks remain visually consistent with the expert labels in crowded regions, elongated structures, and dense erythrocyte fields.

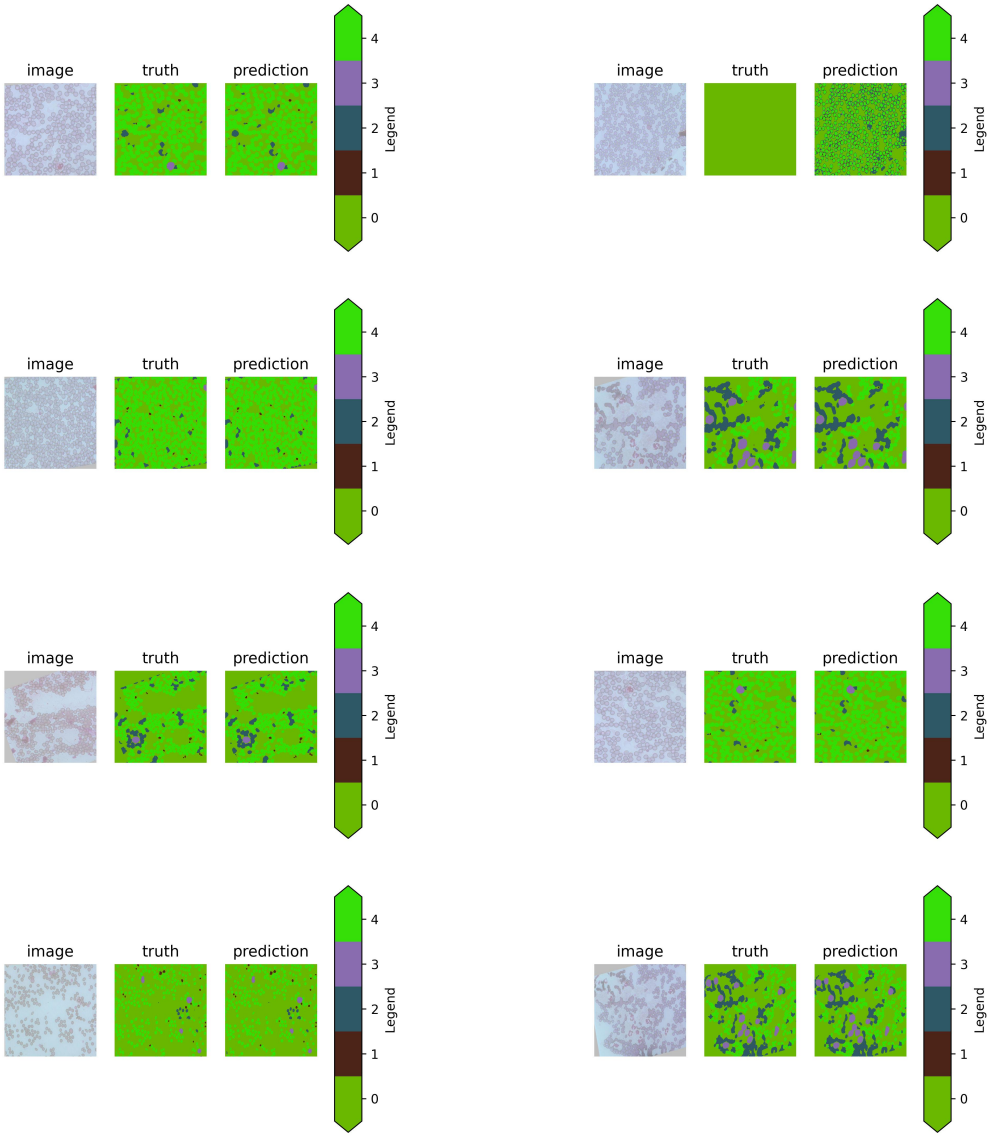


Figure 19: Third set of qualitative comparison examples for the semi-supervised segmentation framework.

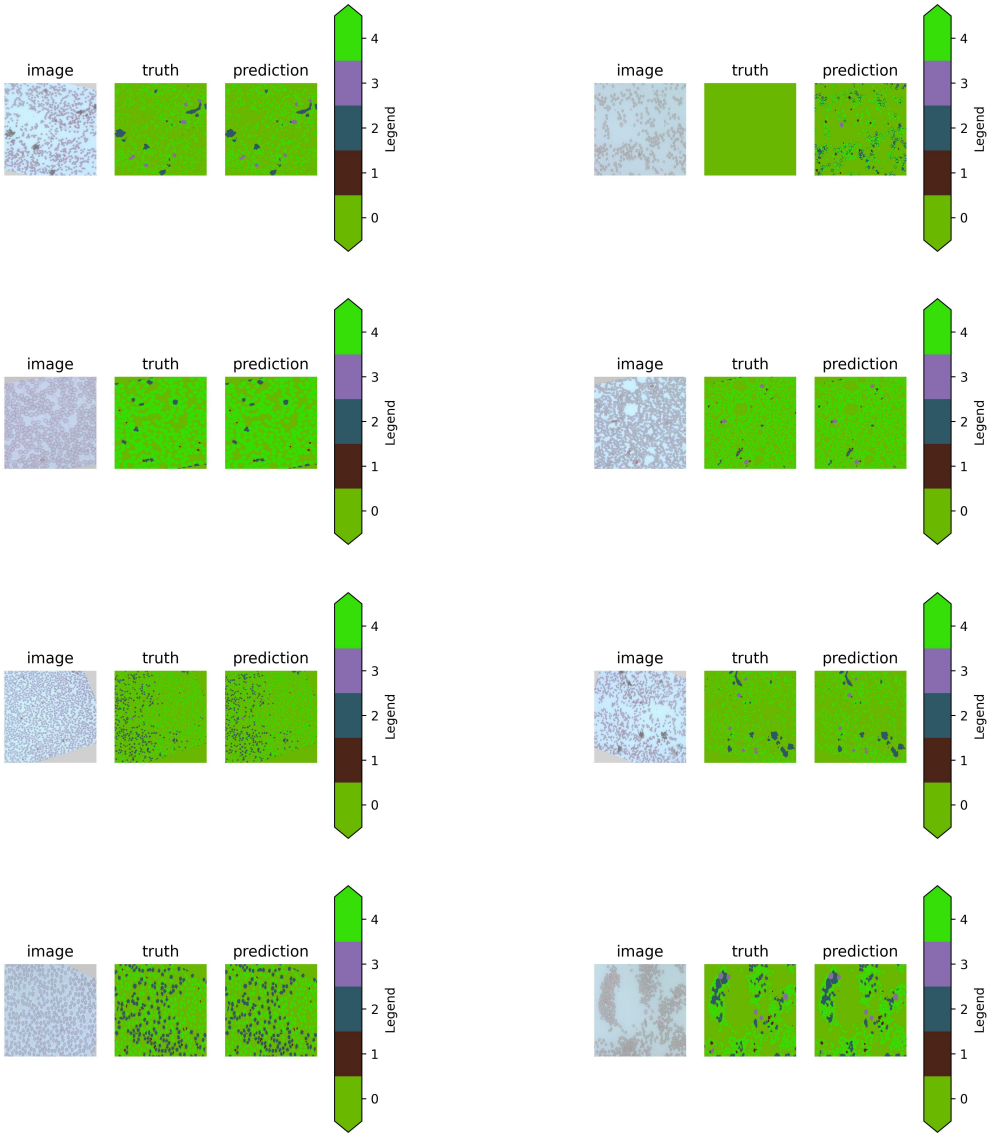


Figure 20: Fourth set of qualitative comparison examples for the semi-supervised segmentation framework.

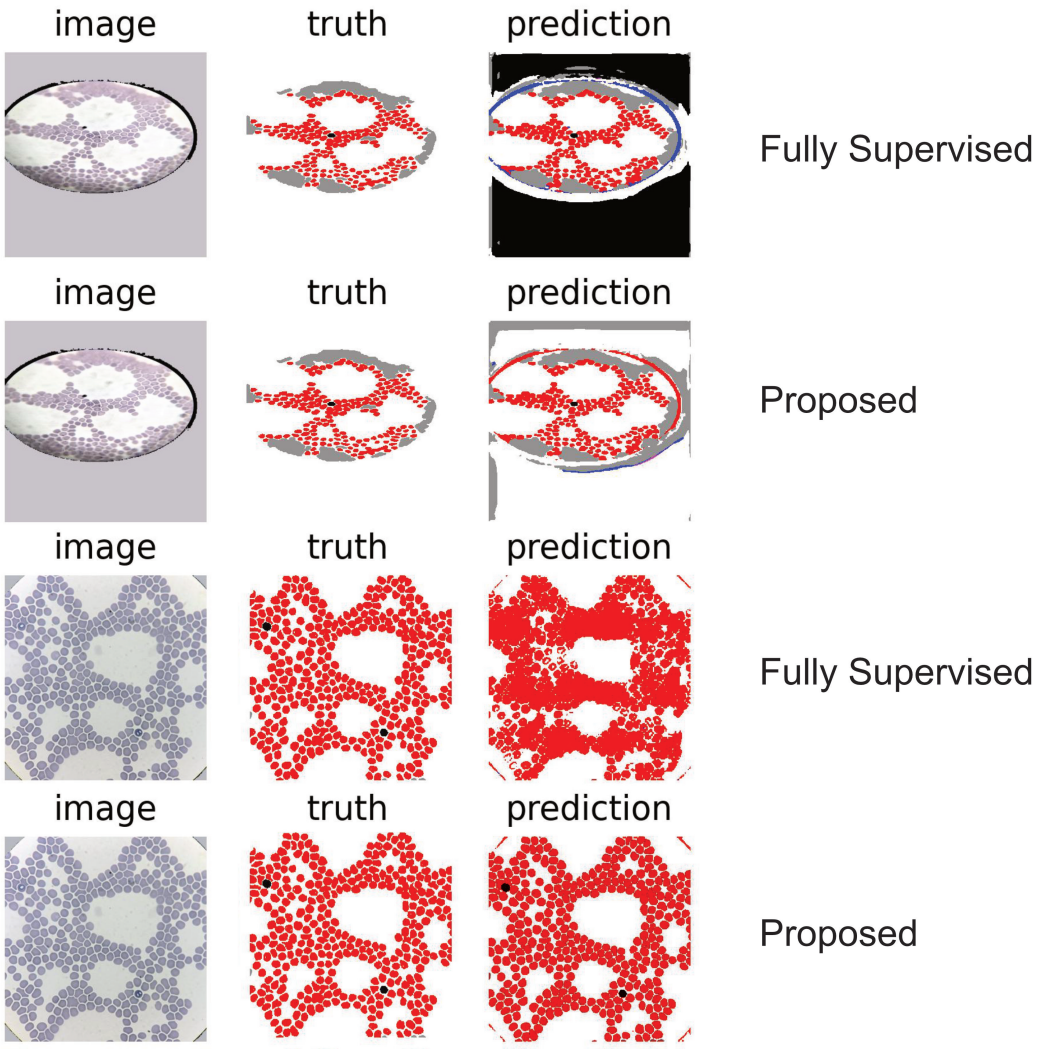


Figure 21: Examples of differences of proposed semi-supervised solution and training on a smaller, fully labeled subset.

10 Integrated discussion

In this integrated discussion, I interpret the preceding chapters as one cumulative research trajectory. I began by showing that task-specific segmentation architectures and Transformer-based context modeling could improve dense medical prediction. I then examined synthetic data generation, fuzzy uncertainty modeling, and additional microscopy tasks. Each of these studies isolated a specific obstacle that had to be resolved before a practical semi-supervised Transformer for hematological microscopy could be formulated.

10.1 The self-attention bottleneck in high-resolution microscopy

The computational limitation of standard self-attention was already formalized earlier in Sections 3.6–3.8. In high-resolution blood-smear microscopy, the practical consequence is that small patches are needed to preserve boundary detail, but that same choice rapidly increases the token count and makes dense attention prohibitively expensive in both memory and runtime. This is the main bottleneck that motivated the pooled-memory compression stage of the final model.

The pooled-memory alternative. A common response to this difficulty is to restrict attention to local windows or hierarchical neighbourhoods. Those strategies reduce cost, but they also weaken long-range interactions that may still be useful when stain balance or morphological context must be interpreted across a broad field of view.

My alternative is to compress the key and value sequences into a compact pooled memory bank. The model attends from all query positions to a carefully constructed set of pooled context tokens, enriched by both interpolation-based context summaries and one global average token. This reduces cost substantially while preserving the components that matter most for cell boundaries, stain balance, and field-wide morphology.

10.2 Morphological variability and staining artifacts

The characteristics of the dataset further motivate the architectural choices. Romanowsky staining introduces variation in hue and contrast across clinics and across samples. In addition, biological morphology varies between species and between cell types. Feline platelet clusters and canine stippling patterns are examples of structures that can be difficult to interpret from local evidence alone.

A model that relies only on short-range convolutional context may confuse these patterns with artifacts or merge neighbouring structures incorrectly. In the proposed

framework, global attention, spectral filtering, and local refinement through MixFFN are intended to work together: attention captures wider context, spectral filtering reduces redundant low-frequency variation, and the local block preserves the spatial detail needed for cell boundaries.

10.3 Extending fuzzy logic to deep semantic maps

The basic terminology of fuzzy sets, memberships, and FCM was introduced earlier in Section 2. In the present framework, the relevant consequence is practical rather than definitional: ambiguous cell borders are better represented by partial membership than by forced one-hot assignment.

Within-cluster objective. The RFCM+ term introduced earlier in this chapter uses the fuzzifier m to control how sharp or diffuse those memberships remain. In validation experiments, values near $m \approx 1.5$ provided a useful balance between continuity and class separation. By estimating class prototypes from the input intensities and current predictions, the objective also allows unlabeled data to organize itself into morphologically meaningful groups rather than remaining unused.

Spatial regularization against noise. Standard FCM does not by itself enforce spatial coherence. As a result, isolated dust particles or noise spikes can be grouped incorrectly with biologically meaningful structures if their intensity happens to be similar. The spatial regularizer addresses this by penalizing local disagreement between a pixel's membership and the memberships observed in its neighbourhood.

The regularizer therefore suppresses isolated high-frequency noise and encourages contiguous segmentations that better match the expected geometry of cells and background regions.

10.4 Architectural synthesis of the training workflow

These components become useful only when they are combined within a single training process. The unsupervised RFCM+ term organizes unlabeled images into coherent latent groups, while the masked supervised term attaches those groups to expert-defined classes whenever annotations are available. The overall system therefore uses sparse supervision as an anchor rather than as the sole source of learning.

10.5 Hardware considerations and inference optimization

Even with spectral compression, the model remains computationally demanding because high-resolution microscopy produces large intermediate representations. Prac-

tical training therefore required explicit attention to memory usage and optimization stability.

Mixed-precision arithmetic and memory management. I used Automated Mixed Precision (AMP) to reduce memory consumption during training. By storing selected activations and operations in FP16 while preserving numerically sensitive steps where necessary, this strategy reduced peak memory relative to a full FP32 pipeline.

Because the fuzzy losses involve small fractional values, loss scaling was used to reduce the risk of numerical underflow during backpropagation.

Learning-rate scheduling and convergence. Training stability also depended on the optimization schedule. I used Adam together with a validation-driven `ReduceLROnPlateau` policy rather than a fixed decay schedule. This choice is more appropriate for the semi-supervised setting because the validation signal can stagnate or improve irregularly as the balance between masked supervision and RFCM+ regularization evolves during training. Reducing the learning rate only when validation loss plateaus proved more compatible with the pooled-memory attention blocks and the fuzzy losses than forcing a predetermined schedule.

10.6 Statistical results and ablation studies

The evaluation focused on mean Intersection over Union (mIoU), Pixel Accuracy (PA), and Dice Similarity Coefficient (DSC). The best results were obtained when the unsupervised RFCM+ term was combined with the masked supervised term, indicating that the two objectives are complementary rather than redundant.

In the feline validation experiments, the semi-supervised system remained competitive even when only a small fraction of the masks were available for direct supervision. This suggests that the unlabeled data were being used constructively rather than merely regularizing an otherwise supervised solution.

The ablation studies also clarified the role of spectral compression. Removing the FFT stage restored the quadratic cost of standard attention and substantially increased training time on the same hardware. Retaining the compression stage reduced resource usage while preserving strong segmentation quality, which supports the practical motivation for the proposed design.

11 Conclusion

This dissertation addressed one central scientific problem: how to design AI systems for medical and biologically relevant image analysis that remain accurate, data-efficient, computationally feasible, and reliable in the presence of uncertainty. I structured the research as a sequence of linked experimental studies in which each chapter isolated one obstacle that prevented the direct construction of the final hematological segmentation framework. Those obstacles included architectural inadequacy, ambiguity of labels and boundaries, and scarcity of dense annotations.

The early chapters established that custom neural architectures can outperform generic designs in demanding medical tasks. The middle chapters showed that synthetic data, dense prediction, and controlled data augmentation can each solve a specific practical bottleneck. The later chapters introduced fuzzy uncertainty modeling as a formal component of learning and then integrated it with Transformer-based segmentation in hematological microscopy. The final framework is therefore not an isolated result. It is the methodological synthesis of the entire dissertation.

11.1 Verification of the dissertation theses

The three theses are verified jointly by the cumulative structure of the dissertation. Some chapters provide direct empirical verification in the target setting of hematological segmentation, whereas others validate the enabling principles that are later incorporated into the final framework.

The first thesis stated that the development of new latent-compression attention mechanisms for image processing can improve the generalization capabilities of Transformer-based semantic segmentation for animal blood sample analysis. The dissertation validates this thesis cumulatively, with the final hematological study providing the direct task-level verification. First, the earlier Transformer-based microscopy study showed that global-context modeling is effective for fine biological structures: the proposed TSSN achieved **98.45%** test accuracy, **0.917** mIoU, and **0.955** mDice, with **0.842** IoU on the rare spirochaete class. This result established the supporting architectural premise that Transformer-style context aggregation is beneficial for thin, visually difficult structures that cannot be resolved from local convolution alone. Second, the final hematological framework introduced a stronger architectural mechanism directly in blood-sample segmentation: pooled-memory global attention combined with spectral compression of key and value streams. In the final validation protocol, this model achieved **99.73%** validation accuracy, **99.73%** micro-Dice, and **99.46%** micro-IoU in a regime with **1200** microscopy images but only **80** expert masks. The ablation analysis further showed that removing the FFT-based compression restored the practical cost of dense attention without delivering a

corresponding benefit in segmentation quality. Thesis 1 is therefore confirmed within the target application domain: compressed attention is not only computationally justified, but also scientifically effective for hematological semantic segmentation.

The second thesis stated that a fuzzy-based semi-supervised training algorithm for medical image segmentation can improve pixel classification under ambiguous data conditions. The direct verification of this claim comes from the final hematological study. The framework was designed specifically for a mixed labeled-unlabeled regime, and the best results were obtained when the unsupervised RFCM+ term was combined with the masked supervised branch rather than when either component was used in isolation. In other words, unlabeled morphology was not merely acting as weak regularization; it improved the quality of the learned representation and the resulting pixel assignments. The model also remained competitive when only a small fraction of the dataset was directly annotated, which is precisely the data regime that motivates semi-supervised learning in hematological practice. Additional chapters reinforce the same dissertation-level claim from complementary directions. The synthetic-data segmentation study achieved **97.3%** IoU and **0.9932** F1, while training on a small real dataset yielded only **85.6%** IoU, demonstrating quantitatively that performance in data-scarce domains improves when learning is allowed to exploit more than direct manual labeling. Thesis 2 is therefore confirmed: a semi-supervised learning strategy grounded in fuzzy clustering and sparse supervision improves pixel classification under annotation scarcity and ambiguity.

The third thesis stated that the introduction of fuzzy logic into semantic-segmentation architectures can reduce pixel-level classification errors on ambiguous data and object borders. This claim is supported both methodologically and directly. First, the dissertation established fuzzy logic as an operational rather than merely descriptive formalism. In the federated-learning chapter, the proposed Type-2 Fuzzy Reputation mechanism preserved **72.6%** global accuracy under the most difficult adaptive attack, while reducing false rejection rate to **8.3%**, false acceptance rate to **11.7%**, detection delay to **1.8** rounds, and improving recovery to **91.2%**. These results demonstrate, at the methodological level, that fuzzy inference improves decision quality when evidence is uncertain, overlapping, and non-crisp. Second, in the final hematological segmentation framework, fuzzy membership modeling was transferred directly to the pixel level through RFCM+ regularization and masked fuzzy supervision. The strongest gains were observed in precisely the regions that motivated the thesis statement: weakly stained borders, overlapping erythrocytes, platelet aggregates, and crowded smear fields where a hard one-hot label is often an inadequate description of the image evidence. The fact that the best overall results were obtained only when the fuzzy term was retained is direct experimental confirmation that fuzzy logic reduces the error introduced by forcing premature crisp

decisions in ambiguous regions. Thesis 3 is therefore confirmed.

Taken together, these results make the logical status of the dissertation explicit. The earlier chapters establish the necessary architectural, data-efficiency, and uncertainty-handling premises, and the final hematological framework verifies their joint effectiveness in the target domain. Accordingly, the material presented in the dissertation is sufficient to treat all three stated theses as proved within the scope of the reported experiments, datasets, and evaluation protocols.

11.2 Scientific contributions of the dissertation

The dissertation makes a scientific contribution at the level of architectural design. The TSSN study showed that Transformer-based context modeling can be adapted to microscopy when paired with a decoder that restores spatial detail, reaching strong segmentation performance on fine biological structures. The later gastrointestinal and malaria studies reinforced the same principle from complementary perspectives: architecture should be derived from the structure of the problem rather than adopted as a generic template. Across these chapters, the dissertation demonstrates that strong results arise when representation learning, decoder design, preprocessing, and optimization are matched to the imaging task itself.

The dissertation also contributes a coherent treatment of data economy. In the synthetic-data chapter, I showed that dense segmentation can be trained successfully even when real annotations are scarce or impractical to collect. The weapon-segmentation system reached **99.6%** pixel accuracy, **97.3%** weapon IoU, and **0.9932** image-level F1, and the ablation study demonstrated that synthetic-first training was decisively stronger than a small manually annotated real-data alternative. In the gastrointestinal histology chapter, mathematically controlled augmentation increased stability and accuracy, producing **98.34%** validation top-1 accuracy, **98.30** Cohen's kappa, and stable cross-validation performance with mean top-1 accuracy of **98.12%** and standard deviation of only **0.23%**. Together, these results show that high-quality AI in small-data scientific domains depends on principled data strategy as much as on network depth.

Another contribution of the dissertation is the integration of dense prediction into biologically meaningful image understanding. The TSSN chapter demonstrated that Transformer-based segmentation can work effectively in microscopy when combined with a decoder that restores spatial detail. The malaria chapter moved closer to the final target domain of this thesis and showed that a custom segmentation network can support blood-smear analysis with both spatial precision and diagnostic utility. The model achieved **97.1%** per-pixel segmentation accuracy and **99.68%** cell-level diagnostic accuracy, outperforming alternative detection-oriented methods.

These chapters were preparatory experiments that clarified how global context, skip-connected reconstruction, preprocessing, and dense supervision should be organized before attempting the final hematological synthesis.

The dissertation further contributes a rigorous treatment of uncertainty. The federated-learning study showed that fuzzy logic can improve robustness in systems where uncertainty exists not in pixels but in behavior and trust. This result broadens the methodological scope of the thesis by showing that fuzzy inference remains valuable when evidence is overlapping, incomplete, and operationally noisy, which strengthens the rationale for introducing fuzzy membership modeling into the final hematological segmentation framework.

The principal contribution of the dissertation is the final semi-supervised Transformer framework for hematological semantic segmentation. This model integrates stain-aware normalization, pooled-memory attention, spectral compression, mixed labeled-unlabeled batches, masked supervised optimization, and RFCM+-based fuzzy regularization within a single training process. Scientifically, the importance of this framework is that it addresses the true structure of the blood-smear problem rather than an idealized simplified version of it. Blood-smear images are not only high-dimensional; they are ambiguous, imbalanced, weakly annotated, morphologically heterogeneous, and often difficult even for specialists. The final model responds to these properties directly. Its best validation performance of **99.73%** accuracy, **99.73%** micro-Dice, and **99.46%** micro-IoU, obtained in a low-label regime, is therefore not merely a strong numerical result. It is experimental evidence that the central design principles of the dissertation operate coherently when combined in one system.

11.3 Final assessment and future research

The results of the dissertation support one general scientific conclusion: high-quality semantic segmentation for medical and veterinary image analysis cannot be obtained reliably by optimizing architecture, supervision, data efficiency, and uncertainty handling in isolation. These dimensions interact. A model may be expressive but not trainable; accurate on abundant labels but ineffective in sparse-label regimes; efficient in memory and compute yet unable to preserve biologically important boundaries; or numerically strong on clean benchmarks while failing in ambiguous clinical imagery. The research presented here shows that these trade-offs can be resolved more effectively when the system is treated as an integrated scientific instrument rather than as a disconnected collection of components.

The dissertation also identifies the limits of the present work with scientific clarity. The current framework is semantic rather than instance-based, which means that

object counting and explicit separation of touching structures still require additional post-processing. More generally, large-scale external validation across laboratories, species, staining protocols, microscopes, and acquisition pipelines remains necessary before any clinical deployment claim could be considered complete. These are not weaknesses of the central thesis; they are the natural next research questions opened by the present results.

Future work should therefore proceed along several concrete directions. First, the final hematological framework should be validated prospectively on new multi-center datasets and extended with stronger per-class overlap analysis, calibration analysis, and expert-reader studies. Second, the current semantic formulation should be expanded toward instance-aware or panoptic blood-smear analysis so that segmentation, counting, and morphology assessment can be performed jointly. Third, active learning and uncertainty-guided annotation should be integrated with the fuzzy semi-supervised objective to reduce expert labeling burden even further. Fourth, model compression, quantization, and distillation should be applied to the final Transformer so that the benefits of the method can be transferred to lower-power clinical and veterinary hardware. Finally, the relationship between fuzzy memberships and probabilistic calibration deserves deeper theoretical study, because this may provide a more unified account of uncertainty in medical AI than either crisp pseudo-labeling or purely deterministic segmentation losses alone.

In conclusion, this dissertation demonstrates that the route to effective AI for hematological microscopy is not a single architectural novelty, but a research program that combines custom representation learning, principled data efficiency, dense prediction design, and explicit treatment of ambiguity. The final semi-supervised Transformer framework confirms that these elements can be brought together successfully. For that reason, the dissertation does more than present a high-performing model. It provides an experimentally grounded methodological framework for future work on trustworthy, efficient, and uncertainty-aware semantic segmentation in medical and veterinary imaging.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. DOI: 10.1038/nature14539. [Online]. Available: https://doi.org/10.1038/nature14539.

[2] G. Litjens et al., "A survey on deep learning in medical image analysis", *Medical Image Analysis*, vol. 42, pp. 60–88, 2017. DOI: 10.1016/j.media.2017.07.005. [Online]. Available: https://doi.org/10.1016/j.media.2017.07.005.

[3] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis", *Annual Review of Biomedical Engineering*, vol. 19, no. 1, pp. 221–248, 2017. DOI: 10.1146/annurev-bioeng-071516-044442. [Online]. Available: https://doi.org/10.1146/annurev-bioeng-071516-044442.

[4] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks", *Nature*, vol. 542, no. 7639, pp. 115–118, 2017. DOI: 10.1038/nature21056. [Online]. Available: https://doi.org/10.1038/nature21056.

[5] E. H. Shortliffe, "A rule-based computer program for advising physicians regarding antimicrobial therapy selection", in *Proceedings of the 1974 Annual ACM Conference - Volume 2*, ser. ACM '74, New York, NY, USA: Association for Computing Machinery, 1974, p. 739, ISBN: 9781450378505. DOI: 10.1145/1408800.1408906. [Online]. Available: https://doi.org/10.1145/1408800.1408906.

[6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. DOI: 10.1109/5.726791. [Online]. Available: https://doi.org/10.1109/5.726791.

[7] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale", in *9th International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy.

[8] A. Vaswani et al., "Attention is all you need", in *Advances in Neural Information Processing Systems 30*, 2017, pp. 5998–6008. [Online]. Available: https://papers.nips.cc/paper/7181-attention-is-all-you-need.

[9] J. Chen et al., "Transunet: Transformers make strong encoders for medical image segmentation", *arXiv preprint arXiv:2102.04306*, 2021. [Online]. Available: https://arxiv.org/abs/2102.04306.

[10] H. Xiao, L. Li, Q. Liu, X. Zhu, and Q. Zhang, "Transformers in medical image segmentation: A review", *Biomedical Signal Processing and Control*, vol. 84, p. 104 791, 2023. DOI: 10.1016/j.bspc.2023.104791. [Online]. Available: https://doi.org/10.1016/j.bspc.2023.104791.

[11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440. DOI: 10.1109/CVPR.2015.7298965.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation", in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

[13] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation", in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. [Online]. Available: https://arxiv.org/abs/1802.02611.

[14] A. Hatamizadeh et al., "UNETR: Transformers for 3d medical image segmentation", in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 1748–1758. DOI: 10.1109/WACV51458.2022.00181. [Online]. Available: https://doi.org/10.1109/WACV51458.2022.00181.

[15] H. Cao et al., "Swin-unet: Unet-like pure transformer for medical image segmentation", in *Computer Vision – ECCV 2022 Workshops*, Springer Nature Switzerland, 2023, pp. 205–218. DOI: 10.1007/978-3-031-25066-8_9.

[16] O. N. Manzari, J. M. Kaleybar, H. Saadat, and S. Maleki, "Befunet: A hybrid cnn-transformer architecture for precise medical image segmentation", *arXiv preprint arXiv:2402.08793*, 2024. DOI: 10.48550/arXiv.2402.08793. [Online]. Available: https://doi.org/10.48550/arXiv.2402.08793.

[17] L. A. Zadeh, "Fuzzy sets", *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965. DOI: 10.1016/S0019-9958(65)90241-X. [Online]. Available: https://doi.org/10.1016/S0019-9958(65)90241-X.

[18] E. H. Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller", *International Journal of Man-Machine Studies*, vol. 7, no. 1, pp. 1–13, 1975. DOI: 10.1016/S0020-7373(75)80002-2. [Online]. Available: https://doi.org/10.1016/S0020-7373(75)80002-2.

[19] J. M. Mendel and R. I. B. John, "Type-2 fuzzy sets made simple", *IEEE Transactions on Fuzzy Systems*, vol. 10, no. 2, pp. 117–127, 2002. DOI: 10.1109/91.995115. [Online]. Available: https://doi.org/10.1109/91.995115.

[20] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm", *Computers & Geosciences*, vol. 10, no. 2–3, pp. 191–203, 1984. DOI: 10.1016/0098-3004(84)90020-7. [Online]. Available: https://doi.org/10.1016/0098-3004(84)90020-7.

[21] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks", in *Workshop on Challenges in Representation Learning, ICML*, 2013. [Online]. Available: https://cir.nii.ac.jp/crid/1370013245438200707.

[22] R. Jiao et al., "Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation", *Computers in Biology and Medicine*, vol. 169, p. 107 840, 2024. DOI: 10.1016/j.compbiomed.2023.107840. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2023.107840.

[23] N. M. Deshpande, S. Gite, and R. Aluvalu, "A review of microscopic analysis of blood cells for disease detection with ai perspective", *PeerJ Computer Science*, vol. 7, e460, 2021.

[24] G. Drałus, D. Mazur, and A. Czmil, "Automatic detection and counting of blood cells in smear images using retinanet", *Entropy*, vol. 23, no. 11, p. 1522, 2021. DOI: 10.3390/e23111522. [Online]. Available: https://doi.org/10.3390/e23111522.

[25] S. J. Pan and Q. Yang, "A survey on transfer learning", *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010. DOI: 10.1109/TKDE.2009.191. [Online]. Available: https://doi.org/10.1109/TKDE.2009.191.

[26] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-efficient learning of deep networks from decentralized data", in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 54, 2017, pp. 1273–1282. [Online]. Available: https://proceedings.mlr.press/v54/mcmahan17a.html.

[27] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world", in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017. DOI: 10.1109/IROS.2017.8202133. [Online]. Available: https://doi.org/10.1109/IROS.2017.8202133.

[28] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning", *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019. DOI: 10.1186/s40537-019-0197-0. [Online]. Available: https://doi.org/10.1186/s40537-019-0197-0.

[29] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity", *The Bulletin of Mathematical Biophysics*, vol. 5, no. 4, pp. 115–133, 1943. DOI: 10.1007/BF02478259. [Online]. Available: https://doi.org/10.1007/BF02478259.

[30] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain.", *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[31] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors", *Nature*, vol. 323, no. 6088, pp. 533–536, 1986. DOI: 10.1038/323533a0. [Online]. Available: https://doi.org/10.1038/323533a0.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", in *3rd International Conference on Learning Representations (ICLR)*, 2015. [Online]. Available: https://arxiv.org/abs/1412.6980.

[33] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts", in *5th International Conference on Learning Representations (ICLR)*, 2017. [Online]. Available: https://arxiv.org/abs/1608.03983.

[34] J. L. Elman, "Finding structure in time", *Cognitive Science*, vol. 14, no. 2, pp. 179–211, 1990. DOI: 10.1207/s15516709cog1402_1.

[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735.

[36] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation", in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734. DOI: 10.3115/v1/D14-1179. [Online]. Available: https://aclanthology.org/D14-1179/.

[37] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate", in *3rd International Conference on Learning Representations (ICLR)*, 2015. [Online]. Available: https://arxiv.org/abs/1409.0473.

[38] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520. DOI: 10.1109/CVPR.2018.00474. [Online]. Available: https://doi.org/10.1109/CVPR.2018.00474.

[39] N. Wojtas, M. Wieczorek, and Z. Bełkot, "Malaria detection using custom semantic segmentation neural network architecture", *Medycyna Weterynaryjna*, vol. 79, no. 8, pp. 406–412, 2023. DOI: 10.21521/mw.6804. [Online]. Available: https://doi.org/10.21521/mw.6804.

[40] W. Siłka, M. Wieczorek, J. Siłka, and M. Woźniak, "Malaria detection using advanced deep learning architecture", *Sensors*, vol. 23, no. 3, p. 1501, 2023.

[41] World Health Organization, *World Malaria Report 2025*. Geneva, Switzerland: World Health Organization, 2025. [Online]. Available: https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2025.

[42] F. B. Tek, A. G. Dempster, and I. Kale, "Computer vision for microscopy diagnosis of malaria", *Malaria Journal*, vol. 8, no. 1, p. 153, 2009. DOI: 10.1186/1475-2875-8-153. [Online]. Available: https://doi.org/10.1186/1475-2875-8-153.

[43] M. D. Samuel, B. L. Woodworth, C. T. Atkinson, P. J. Hart, and D. A. LaPointe, "Avian malaria in hawaiian forest birds: Infection and population impacts across species and elevations", *Ecosphere*, vol. 6, no. 6, pp. 1–21, 2015. DOI: 10.1890/ES14-00393.1. [Online]. Available: https://doi.org/10.1890/ES14-00393.1.

[44] P. Ruiz Cuenca et al., "Epidemiology of the zoonotic malaria *Plasmodium knowlesi* in changing landscapes", *Advances in Parasitology*, vol. 113, pp. 225–286, 2021. DOI: 10.1016/bs.apar.2021.08.006. [Online]. Available: https://doi.org/10.1016/bs.apar.2021.08.006.

[45] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement", *arXiv preprint arXiv:1804.02767*, 2018. [Online]. Available: https://arxiv.org/abs/1804.02767.

[46] D. Shah, K. Kawale, M. Shah, S. Randive, and R. Mapari, "Malaria parasite detection using deep learning: (beneficial to humankind)", in *Proceedings of the 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 2020, pp. 984–988.

[47] W. R. W. M. Razin, T. S. Gunawan, M. Kartiwi, and N. M. Yusoff, "Malaria parasite detection and classification using CNN and YOLOv5 architectures", in *Proceedings of the 2022 IEEE 8th International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA)*, Melaka, Malaysia, 2022, pp. 277–281.

[48] A. Alqudah, A. M. Alqudah, and S. Qazan, "Lightweight deep learning for malaria parasite detection using cell-image of blood smear images", *Revue d'Intelligence Artificielle*, vol. 34, pp. 571–576, 2020.

[49] Q. Quan, J. Wang, and L. Liu, "An effective convolutional neural network for classifying red blood cells in malaria diseases", *Interdisciplinary Sciences: Computational Life Sciences*, vol. 12, pp. 217–225, 2020.

[50] M. Turuk, R. Sreemathy, S. Kadiyala, S. Kotecha, and V. Kulkarni, "CNN based deep learning approach for automatic malaria parasite detection", *IAENG International Journal of Computer Science*, vol. 49, pp. 745–753, 2022.

[51] G. Shekar, S. Revathy, and E. K. Goud, "Malaria detection using deep learning", in *Proceedings of the 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, 2020, pp. 746–750.

[52] A. Rahman et al., "Improving malaria parasite detection from red blood cell using deep convolutional neural networks", *arXiv preprint arXiv:1907.10418*, 2019. [Online]. Available: https://arxiv.org/abs/1907.10418.

[53] D. R. Loh, W. X. Yong, J. Yapeter, K. Subburaj, and R. Chandramohanadas, "A deep learning approach to the screening of malaria infection: Automated and rapid cell counting, object detection and instance segmentation using Mask R-CNN", *Computerized Medical Imaging and Graphics*, vol. 88, p. 101 845, 2021. DOI: 10.1016/j.compmedimag.2020.101845. [Online]. Available: https://doi.org/10.1016/j.compmedimag.2020.101845.

[54] S. Saglam, F. Tat, and S. Bayar, "FPGA implementation of CNN algorithm for detecting malaria diseased blood cells", in *Proceedings of the 2019 International Symposium on Advanced Electrical and Communication Technologies (ISAECT)*, Rome, Italy, 2019, pp. 1–5.

[55] P. Pongpanitanont, N. Suttidate, M. Nuinoon, N. Khampeeramao, S. Laymanivong, and P. Janwan, "Automated malaria ring form classification in blood smear images using ensemble parallel neural networks", *Journal of Imaging*, vol. 12, no. 3, p. 127, 2026. DOI: 10.3390/jimaging12030127. [Online]. Available: https://doi.org/10.3390/jimaging12030127.

[56] R. Parveen et al., "Trustworthy deep learning for malaria diagnosis using explainable artificial intelligence", *Scientific Reports*, vol. 15, p. 45 037, 2025. DOI: 10.1038/s41598-025-28387-7. [Online]. Available: https://doi.org/10.1038/s41598-025-28387-7.

[57] O. R. Shahin, H. H. Alshammari, R. N. Alabdali, A. M. Salaheldin, and N. Saleh, "Automated multi-model framework for malaria detection using deep learning and feature fusion", *Scientific Reports*, vol. 15, p. 25 672, 2025. DOI: 10.1038/s41598-025-04784-w. [Online]. Available: https://doi.org/10.1038/s41598-025-04784-w.

[58] M. Wieczorek, J. Siłka, K. Wiltos, and M. Woźniak, "Transformer based semantic segmentation network for medical imaging application", in *International Conference on Artificial Intelligence and Soft Computing*, Springer, 2024, pp. 380–389.

[59] N. Otsu, "A threshold selection method from gray-level histograms", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979. DOI: 10.1109/TSMC.1979.4310076. [Online]. Available: https://doi.org/10.1109/TSMC.1979.4310076.

[60] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models", *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988. DOI: 10.1007/BF00133570. [Online]. Available: https://doi.org/10.1007/BF00133570.

[61] S. Osher and J. A. Sethian, "Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations", *Journal of Computational Physics*, vol. 79, no. 1, pp. 12–49, 1988. DOI: 10.1016/0021-9991(88)90002-2. [Online]. Available: https://doi.org/10.1016/0021-9991(88)90002-2.

[62] M. Cabezas, A. Oliver, X. Lladó, J. Freixenet, and M. B. Cuadra, "A review of atlas-based segmentation for magnetic resonance brain images", *Computer Methods and Programs in Biomedicine*, vol. 104, no. 3, e158–e177, 2011. DOI: 10.1016/j.cmpb.2011.07.015. [Online]. Available: https://doi.org/10.1016/j.cmpb.2011.07.015.

[63] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation", in *2016 Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 565–571. DOI: 10.1109/3DV.2016.79. [Online]. Available: https://doi.org/10.1109/3DV.2016.79.

[64] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017. DOI: 10.1109/TPAMI.2016.2644615. [Online]. Available: https://doi.org/10.1109/TPAMI.2016.2644615.

[65] O. Oktay et al., "Attention u-net: Learning where to look for the pancreas", *arXiv preprint arXiv:1804.03999*, 2018. DOI: 10.48550/arXiv.1804.03999. [Online]. Available: https://arxiv.org/abs/1804.03999.

[66] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation", in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2018, pp. 3–11. DOI: 10.1007/978-3-030-00889-5_1.

[67] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "Doubleu-net: A deep convolutional neural network for medical image segmentation", in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, 2020, pp. 558–564. DOI: 10.1109/CBMS49503.2020.00111. [Online]. Available: https://doi.org/10.1109/CBMS49503.2020.00111.

[68] A. G. Roy, N. Navab, and C. Wachinger, "Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks", *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 540–549, 2019. DOI: 10.1109/TMI.2018.2867261. [Online]. Available: https://doi.org/10.1109/TMI.2018.2867261.

[69] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "Nnu-net: A self-configuring method for deep learning-based biomedical image segmentation", *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021. DOI: 10.1038/s41592-020-01008-z. [Online]. Available: https://doi.org/10.1038/s41592-020-01008-z.

[70] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7794–7803. DOI: 10.1109/CVPR.2018.00813. [Online]. Available: https://doi.org/10.1109/CVPR.2018.00813.

[71] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6881–6890. DOI: 10.1109/CVPR46437.2021.00681. [Online]. Available: https://doi.org/10.1109/CVPR46437.2021.00681.

[72] Y. Xu, S. Hou, X. Wang, D. Li, and L. Lu, "A medical image segmentation method based on improved unet 3+ network", *Diagnostics*, vol. 13, no. 3, p. 576, 2023. DOI: 10.3390/diagnostics13030576. [Online]. Available: https://doi.org/10.3390/diagnostics13030576.

[73] A. Beji, A. G. Blaiech, M. Said, A. Ben Abdallah, and M. H. Bedoui, "An innovative medical image synthesis based on dual gan deep neural networks for improved segmentation quality", *Applied Intelligence*, vol. 53, no. 3, pp. 3381–3397, 2023. DOI: 10.1007/s10489-022-03682-2. [Online]. Available: https://doi.org/10.1007/s10489-022-03682-2.

[74] X. Du, Y. Nie, F. Wang, T. Lei, S. Wang, and X. Zhang, "Al-net: Asymmetric lightweight network for medical image segmentation", *Frontiers in Signal Processing*, vol. 2, p. 842 925, 2022. DOI: 10.3389/frsip.2022.842925. [Online]. Available: https://doi.org/10.3389/frsip.2022.842925.

[75] Y. Wu, Y. Xia, Y. Song, Y. Zhang, and W. Cai, "Nfn+: A novel network followed network for retinal vessel segmentation", *Neural Networks*, vol. 126, pp. 153–162, 2020. DOI: 10.1016/j.neunet.2020.02.018. [Online]. Available: https://doi.org/10.1016/j.neunet.2020.02.018.

[76] Y. Xu, Y. Wang, J. Yuan, Q. Cheng, X. Wang, and P. L. Carson, "Medical breast ultrasound image segmentation by machine learning", *Ultrasonics*, vol. 91, pp. 1–9, 2019. DOI: 10.1016/j.ultras.2018.07.006. [Online]. Available: https://doi.org/10.1016/j.ultras.2018.07.006.

[77] M. Wieczorek, J. Siłka, M. Kobielnik, and M. Woźniak, "Semantic segmentation neural network in automatic weapon detection", in *International Conference on Artificial Intelligence and Soft Computing*, Springer, 2023, pp. 112–120.

[78] R. Olmos, S. Tabik, and F. Herrera, "Automatic handgun detection alarm in videos using deep learning", *Neurocomputing*, vol. 275, pp. 66–72, 2018. DOI: 10.1016/j.neucom.2017.05.012. [Online]. Available: https://doi.org/10.1016/j.neucom.2017.05.012.

[79] J. Salido, V. Lomas, J. Ruiz-Santaquiteria, and O. Deniz, "Automatic handgun detection with deep learning in video surveillance images", *Applied Sciences*, vol. 11, no. 13, p. 6085, 2021. DOI: 10.3390/app11136085. [Online]. Available: https://doi.org/10.3390/app11136085.

[80] S. I. Nikolenko, *Synthetic Data for Deep Learning*. Cham, Switzerland: Springer, 2021. DOI: 10.1007/978-3-030-75178-4. [Online]. Available: https://doi.org/10.1007/978-3-030-75178-4.

[81] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks", in *Advances in Neural Information Processing Systems 28*, 2015. [Online]. Available: https://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.

[82] J. Ruiz-Santaquiteria, O. Deniz, J. Salido, and G. Bueno, "Combined human pose and appearance for convolutional neural networks based armed person detection in surveillance environments", *IEEE Access*, vol. 9, pp. 123 815–123 826, 2021. DOI: 10.1109/ACCESS.2021.3110335. [Online]. Available: https://doi.org/10.1109/ACCESS.2021.3110335.

[83] M. Grega, A. Matiolański, P. Guzik, and M. Leszczuk, "Automated detection of firearms and knives in a cctv image", *Sensors*, vol. 16, no. 1, p. 47, 2016. DOI: 10.3390/s16010047. [Online]. Available: https://doi.org/10.3390/s16010047.

[84] M. T. Bhatti, M. G. Khan, M. Aslam, and M. J. Fiaz, "Weapon detection in real-time cctv videos using deep learning", *IEEE Access*, vol. 9, pp. 34 366–34 382, 2021. DOI: 10.1109/ACCESS.2021.3059170. [Online]. Available: https://doi.org/10.1109/ACCESS.2021.3059170.

[85] P. Yadav, N. Gupta, and P. K. Sharma, "Robust weapon detection in dark environments using yolov7-darkvision", *Digital Signal Processing*, vol. 145, p. 104 342, 2024. DOI: 10.1016/j.dsp.2023.104342. [Online]. Available: https://doi.org/10.1016/j.dsp.2023.104342.

[86] J. D. Muñoz, J. Ruiz-Santaquiteria, O. Deniz, and G. Bueno, "Concealed weapon detection using thermal cameras", *Journal of Imaging*, vol. 11, no. 3, p. 72, 2025. DOI: 10.3390/jimaging11030072. [Online]. Available: https://doi.org/10.3390/jimaging11030072.

[87] J. Tremblay et al., "Training deep networks with synthetic data: Bridging the reality gap by domain randomization", in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018,

pp. 969–977. DOI: 10.1109/CVPRW.2018.00143. [Online]. Available: https://doi.org/10.1109/CVPRW.2018.00143.

[88] H. Jain, A. Vikram, A. Kashyap, and A. Jain, "Weapon detection using artificial intelligence and deep learning for security applications", in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, IEEE, 2020, pp. 193–198.

[89] G. K. Verma and A. Dhillon, "A handheld gun detection using faster r-cnn deep learning", in *Proceedings of the 7th International Conference on Computer and Communication Technology*, 2017, pp. 84–88.

[90] F. Gelana and A. Yadav, "Firearm detection from surveillance cameras using image processing and machine learning techniques", in *Smart Innovations in Communication and Computational Sciences*, ser. Advances in Intelligent Systems and Computing, vol. 851, Springer, Singapore, 2019, pp. 25–34. DOI: 10.1007/978-981-13-2414-7_3.

[91] R. Olmos, S. Tabik, A. Lamas, F. Pérez-Hernández, and F. Herrera, "A binocular image fusion approach for minimizing false positives in handgun detection with deep learning", *Information Fusion*, vol. 49, pp. 271–280, 2019.

[92] R. Debnath and M. K. Bhowmik, "Automatic visual gun detection carried by a moving person", in *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, IEEE, 2020, pp. 208–213.

[93] A. Jaszcz, D. Połap, A. Zielonka, and M. Wieczorek, "Type-2 fuzzy controller as a client reputation mechanism for federated learning", in *2025 IEEE International Conference on Fuzzy Systems (FUZZ)*, IEEE, 2025, pp. 1–4.

[94] P. Kairouz, H. B. McMahan, B. Avent, et al., "Advances and open problems in federated learning", *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021. DOI: 10.1561/2200000083. [Online]. Available: https://doi.org/10.1561/2200000083.

[95] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning", *Future Generation Computer Systems*, vol. 115, pp. 619–640, 2021. DOI: 10.1016/j.future.2020.10.007. [Online]. Available: https://doi.org/10.1016/j.future.2020.10.007.

[96] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent", in *Advances in Neural Information Processing Systems 30*, 2017. [Online]. Available: https://papers.nips.cc/paper/6617-machine-learning-with-adversaries-byzantine-tolerant-gradient-descent.

[97] A. Krizhevsky, "Learning multiple layers of features from tiny images", University of Toronto, Tech. Rep., 2009. [Online]. Available: https://www.cs.toronto.edu/kriz/learning-features-2009-TR.pdf.

[98] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms", *arXiv preprint arXiv:1708.07747*, 2017. [Online]. Available: https://arxiv.org/abs/1708.07747.

[99] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates", in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80, 2018, pp. 5650–5659. [Online]. Available: https://proceedings.mlr.press/v80/yin18a.html.

[100] X. Cao, M. Fang, J. Liu, and N. Z. Gong, "FLTrust: Byzantine-robust federated learning via trust bootstrapping", in *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2021. DOI: 10.14722/ndss.2021.24434. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/fltrust-byzantine-robust-federated-learning-via-trust-bootstrapping/.

[101] S. Shen, S. Tople, and P. Saxena, "AUROR: Defending against poisoning attacks in collaborative deep learning systems", in *Proceedings of the 32nd Annual Conference on Computer Security Applications*, 2016, pp. 508–519. DOI: 10.1145/2991079.2991125. [Online]. Available: https://www.microsoft.com/en-us/research/publication/auror-defending-against-poisoning-attacks-in-collaborative-deep-learning-systems-2/.

[102] M. Wieczorek, N. Wojtas, R. Wituła, A. Krawczyk, and K. Rycerz, "A custom deep learning architecture with image augmentation for intelligent gastrointestinal tract tissue classification", *International Journal of Applied Mathematics and Computer Science*, vol. 34, no. 4, 2024.

[103] M. Wieczorek, N. Wojtas, M. Woźniak, A. Krawczyk, and K. Rycerz, "Transformer based neural network for intelligent classification of selected animals organs", *Neural Computing and Applications*, vol. 38, no. 4, p. 72, 2026.

[104] D. Komura and S. Ishikawa, "Machine learning methods for histopathological image analysis", *Computational and Structural Biotechnology Journal*, vol. 16, pp. 34–42, 2018. DOI: 10.1016/j.csbj.2018.01.001. [Online]. Available: https://doi.org/10.1016/j.csbj.2018.01.001.

[105] J. P. M. Rodriguez et al., "Artificial intelligence as a tool for diagnosis in digital pathology whole slide images: A systematic review", *Journal of Pathology Informatics*, vol. 13, p. 100 138, 2022. DOI: 10.1016/j.jpi.2022.100138. [Online]. Available: https://doi.org/10.1016/j.jpi.2022.100138.

[106] T. Khorshed, M. N. Moustafa, and A. Rafea, "Deep learning for multi-tissue cancer classification of gene expressions (GeneXNet)", *IEEE Access*, vol. 8, pp. 90 615–90 629, 2020. DOI: 10.1109/ACCESS.2020.2992907. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.2992907.

[107] J. Ker, Y. Bai, H. Y. Lee, J. Rao, and L. Wang, "Automated brain histology classification using machine learning", *Journal of Clinical Neuroscience*, vol. 66, pp. 239–245, 2019. DOI: 10.1016/j.jocn.2019.05.019. [Online]. Available: https://doi.org/10.1016/j.jocn.2019.05.019.

[108] S. Zahia, D. Sierra-Sosa, B. Garcia-Zapirain, and A. Elmaghraby, "Tissue classification and segmentation of pressure injuries using convolutional neural networks", *Computer Methods and Programs in Biomedicine*, vol. 159, pp. 51–58, 2018. DOI: 10.1016/j.cmpb.2018.02.018. [Online]. Available: https://doi.org/10.1016/j.cmpb.2018.02.018.

[109] F. Zhang et al., "Liver tissue classification using an auto-context-based deep neural network with a multi-phase training framework", in *Patch-Based Techniques in Medical Imaging*, 2018, pp. 59–66. DOI: 10.1007/978-3-030-00500-9_7.

[110] M. R. McCourt and T. E. Rizzi, "Hematology of dogs", *Schalm's veterinary hematology*, pp. 969–982, 2022.

[111] K. Karthika and S. Poobitha, "Overview on diseases of blood and hematopoietic system in dogs and cats", in *Introduction to Diseases, Diagnosis, and Management of Dogs and Cats*, Elsevier, 2024, pp. 309–320.

[112] S. H. Rezatofighi, K. Khaksari, and H. Soltanian-Zadeh, "Automatic recognition of five types of white blood cells in peripheral blood", in *Image Analysis and Recognition: 7th International Conference, ICIAR 2010, Póvoa de Varzin, Portugal, June 21-23, 2010, Proceedings, Part II 7*, Springer, 2010, pp. 161–172.

[113] A. Woolcock, A. Keenan, C. Cheung, J. Christian, and G. Moore, "Thrombocytosis in 715 dogs (2011–2015)", *Journal of veterinary internal medicine*, vol. 31, no. 6, pp. 1691–1699, 2017.

[114] M. Kraytman, "Platelet size in thrombocytopenias and thrombocytosis of various origin", *Blood*, vol. 41, no. 4, pp. 587–598, 1973.

[115] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers", in *Advances in Neural Information Processing Systems 34*, 2021. [Online]. Available: https://arxiv.org/abs/2105.15203.

[116] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "UCTransNet: Rethinking the skip connections in U-Net from a channel-wise perspective with transformer", in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 2441–2449. DOI: 10.1609/aaai.v36i3.20144. [Online]. Available: https://doi.org/10.1609/aaai.v36i3.20144.

[117] X. Huang, Z. Deng, D. Li, and X. Yuan, "MISSFormer: An effective medical image segmentation transformer", *arXiv preprint arXiv:2109.07162*, 2023. DOI: 10.48550/arXiv.2109.07162. [Online]. Available: https://arxiv.org/abs/2109.07162.

[118] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results", in *5th International Conference on Learning Representations (ICLR), Workshop Track*, 2017. [Online]. Available: https://openreview.net/forum?id=ry8u21rtl.

[119] K. Sohn et al., "Fixmatch: Simplifying semi-supervised learning with consistency and confidence", in *Advances in Neural Information Processing Systems 33*, 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/06964dce9addb1c5cb5d6e3d9838f733-Abstract.html.

[120] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 12 674–12 684. DOI: 10.1109/CVPR42600.2020.01269. [Online]. Available: https://doi.org/10.1109/CVPR42600.2020.01269.

[121] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision", *arXiv preprint arXiv:2106.01226*, 2021. DOI: 10.48550/arXiv.2106.01226. [Online]. Available: https://arxiv.org/abs/2106.01226.

[122] Y. Wang et al., "Semi-supervised semantic segmentation using unreliable pseudo-labels", *arXiv preprint arXiv:2203.03884*, 2022. DOI: 10.48550/arXiv.2203.03884. [Online]. Available: https://arxiv.org/abs/2203.03884.

[123] G. Bortsova, F. Dubost, L. Hogeweg, I. Katramados, and M. de Bruijne, "Semi-supervised medical image segmentation via learning consistency under transformations", in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, ser. Lecture Notes in Computer Science, vol. 11769, Springer, 2019, pp. 810–818. DOI: 10.1007/978-3-030-32226-7_90.

[124] X. Luo, J. Chen, T. Song, and G. Wang, "Semi-supervised medical image segmentation through dual-task consistency", in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 8801–8809. DOI: 10.1609/aaai.v35i10.17066. [Online]. Available: https://doi.org/10.1609/aaai.v35i10.17066.

[125] M. N. Ahmed, S. M. Yamany, N. Mohamed, A. A. Farag, and T. Moriarty, "A modified fuzzy c-means algorithm for bias field estimation and segmentation of mri data", *IEEE Transactions on Medical Imaging*, vol. 21, no. 3, pp. 193–199, 2002. DOI: 10.1109/42.996338. [Online]. Available: https://doi.org/10.1109/42.996338.

[126] S. Krinidis and V. Chatzis, "A robust fuzzy local information c-means clustering algorithm", *IEEE Transactions on Image Processing*, vol. 19, no. 5, pp. 1328–1337, 2010. DOI: 10.1109/TIP.2010.2040763. [Online]. Available: https://doi.org/10.1109/TIP.2010.2040763.

[127] J. Chen et al., "Learning fuzzy clustering for spect/ct segmentation via convolutional neural networks", *Medical Physics*, vol. 48, no. 7, pp. 3860–3877, 2021. DOI: 10.1002/mp.14903. [Online]. Available: https://doi.org/10.1002/mp.14903.

[128] Centers for Disease Control and Prevention, *Clinical overview of babesiosis*, Accessed: 2026-05-09, Feb. 2024. [Online]. Available: https://www.cdc.gov/babesiosis/hcp/clinical-overview/index.html.

[129] R. Welc-Falęciak, A. Pawełczyk, M. Radkowski, S. A. Pancewicz, J. Zajkowska, and E. Siński, "First report of two asymptomatic cases of human infection with *babesia microti* (franca, 1910) in poland", *Annals of Agricultural and Environmental Medicine*, vol. 22, no. 1, pp. 51–54, 2015. DOI: 10.5604/12321966.1141394. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/25780828/.

[130] Centers for Disease Control and Prevention, *DPDx: Babesiosis*, Last reviewed: 2024-06-03; accessed: 2026-05-09, Jun. 2024. [Online]. Available: https://www.cdc.gov/dpdx/babesiosis/index.html.

[131] A. Kloc and E. M. Galińska, "Preliminary assessment of the prevalence of antibodies to *babesia* spp. in a group occupationally exposed to tick bites in poland", *Medycyna Ogólna i Nauki o Zdrowiu*, vol. 29, no. 1, pp. 46–56, 2023. DOI: 10.26444/monz/158846. [Online]. Available: https://www.monz.pl/Wstepna-ocena-czestosci-wystepowania-przeciwcial-anty-Babesia-spp-w-grupie-osob-zawodowo,158846,0,2.html.

[132] D. Dwużnik-Szarek et al., "Monitoring the expansion of *dermacentor reticulatus* and occurrence of canine babesiosis in poland in 2016–2018", *Parasites & Vectors*, vol. 14, p. 267, 2021. DOI: 10.1186/s13071-021-04758-7. [Online]. Available: https://link.springer.com/article/10.1186/s13071-021-04758-7.

[133] D. Dwużnik-Szarek, E. J. Mierzejewska, D. Kiewra, A. Czułowska, A. Robak, and A. Bajer, "Update on prevalence of *babesia canis* and *rickettsia* spp. in adult and juvenile *dermacentor reticulatus* ticks in the area of poland (2016–2018)", *Scientific Reports*, vol. 12, p. 5755, 2022. DOI: 10.1038/s41598-022-09419-y. [Online]. Available: https://www.nature.com/articles/s41598-022-09419-y.

[134] O. Pawełczyk, P. Iwase, B. Wierzba, M. Kretschmer, R. Wojtyczka, and K. Solarz, "A retrospective epidemiological analysis of microscopically detected babesiosis in dogs of southern poland (2018–2022)", *Pathogens*, vol. 13, no. 12, p. 1104, 2024. DOI: 10.3390/pathogens13121104. [Online]. Available: https://www.mdpi.com/2076-0817/13/12/1104.

[135] Y. Gao, M. Zhou, and D. N. Metaxas, "Utnet: A hybrid transformer architecture for medical image segmentation", in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, ser. Lecture Notes in Computer Science, vol. 12903, Springer, 2021, pp. 61–71. DOI: 10.1007/978-3-030-87199-4_6.

[136] H.-Y. Zhou et al., "Nnformer: Volumetric medical image segmentation via a 3d transformer", *IEEE Transactions on Image Processing*, vol. 32, pp. 4036–4045, 2023. DOI: 10.1109/TIP.2023.3293771. [Online]. Available: https://doi.org/10.1109/TIP.2023.3293771.

[137] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images", *arXiv preprint arXiv:2201.01266*, 2022. DOI: 10.48550/arXiv.2201.01266. [Online]. Available: https://arxiv.org/abs/2201.01266.

[138] J. Ma, F. Li, and B. Wang, "U-mamba: Enhancing long-range dependency for biomedical image segmentation", *arXiv preprint arXiv:2401.04722*, 2024. DOI: 10.48550/arXiv.2401.04722. [Online]. Available: https://arxiv.org/abs/2401.04722.

[139] Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu, "Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation", in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, ser. Lecture Notes in Computer Science, vol. 15008, Springer Nature Switzerland, 2024, pp. 578–588. DOI: 10.1007/978-3-031-72111-3_54.

[140] V. Anand, S. Gupta, D. Koundal, W. Y. Alghamdi, and B. M. Alsharbi, "Deep learning-based image annotation for leukocyte segmentation and classification of blood cell morphology", *BMC Medical Imaging*, vol. 24, p. 83, 2024. DOI: 10.1186/s12880-024-01254-z. [Online]. Available: https://doi.org/10.1186/s12880-024-01254-z.

[141] G. Xu, X. Zhang, X. He, and X. Wu, "Levit-unet: Make faster encoders with transformer for medical image segmentation", *arXiv preprint arXiv:2107.08623*, 2021. DOI: 10.48550/arXiv.2107.08623. [Online]. Available: https://arxiv.org/abs/2107.08623.

[142] T. Wang, Z. Huang, J. Wu, Y. Cai, and Z. Li, "Semi-supervised medical image segmentation with co-distribution alignment", *Bioengineering*, vol. 10, no. 7, p. 869, 2023. DOI: 10.3390/bioengineering10070869. [Online]. Available: https://doi.org/10.3390/bioengineering10070869.