

University of Hradec Králové
Faculty of Informatics and Management
Department of Information Technology

Machine Learning for Microscopic Data Processing
Master's thesis

Author: Bc. RICHARD HLUBUČEK
Field of study: Applied Informatics

Supervisor: doc. RNDr. KAMILA ŠTEKEROVÁ, Ph.D., MSc.

Declaration

I declare that I have prepared this diploma thesis independently under the guidance of the thesis supervisor and have listed all sources and literature used. In the case of using artificial intelligence tools, I have stated their purpose and scope of use and verified the accuracy of the information obtained in this way.

.....
Richard Hlubuček
July 19, 2026

Acknowledgements

Mere words cannot express my gratitude to my supervisor, doc. RNDr. Kamila Štekerová, Ph.D., MSc., for the opportunity to collaborate on a scientific project of this scale. I would like to sincerely thank the entire machine learning science team at FIM UHK for the time and guidance as I shared and developed my ideas throughout the project. My only wish is to keep advancing the applied knowledge in all future endeavors.

Furthermore, I want to thank my wife Kateřina H. for her undying patience, care, understanding and benevolence toward the countless hours I invested into my studies, along with research and development of this thesis. I would like to especially thank our expected baby girl, who was my quietest and most sincere companion and the greatest motivation of all. I would like to thank my entire family for understanding and support.

Finally, I would like to dedicate this work to the loving memory of my grandfather Miloš H., who always cherished the time we spent together and who taught me that even the largest problems and architectures can be broken down and solved with simple steps.

Annotation

While artificial neural networks represent the pinnacle of modern image analysis, the scarcity of validated microscopic data remains a leading bottleneck. This work evaluates three prominent object detection architectures on a novel dataset of complex botanical micro-structures sourced from the Department of Pharmacognosy and Pharmaceutical Botany. The overall best empirical results were achieved by the YOLOv10x model, securing an mAP@0.50 of 77.8% and an mAP@0.50:0.95 of 49.2%. Crucially, model performance was significantly improved by consolidating the highly imbalanced training dataset, which increased precision across individual classes. Furthermore, post-hoc Explainable AI (XAI) techniques, including Grad-CAM and spatial heatmaps, were deployed to validate the diagnostic transparency of the highest-performing network. The experiments demonstrate that single-stage algorithms, coupled with rigorous data pruning and feature validation, offer the greatest potential for automated microscopic data analysis.

Keywords

Machine Learning, Object Detection, Microscopic Image Analysis, Pharmacognosy, Explainable AI (XAI), Deep Neural Networks, Data Augmentation, YOLOv10

Anotace

Zpracování mikroskopických dat metodami strojového učení

Ačkoliv umělé neuronové sítě představují vrchol moderní analýzy obrazu, nedostatek validovaných mikroskopických dat zůstává hlavní překážkou. Tato práce hodnotí tři významné architektury pro detekci objektů na nové datové sadě složitých botanických mikrostruktur, získané z Katedry farmakognozie a farmaceutické botaniky. Celkově nejlepších empirických výsledků dosáhl model YOLOv10x s hodnotami mAP@0.5 na úrovni 77,8 % a mAP@0.5:0.95 na úrovni 49,2 %. Zásadní je, že výkonnost se výrazně zlepšila konsolidací silně nevyvážené trénovací datové sady, což zvýšilo přesnost napříč jednotlivými třídami. K ověření diagnostické transparentnosti nejvýkonnější sítě byly navíc nasazeny metody vysvětlitelné umělé inteligence (XAI), včetně technik Grad-CAM a prostorových teplotních map. Experimenty prokazují, že jednofázové algoritmy ve spojení s důslednou redukcí dat a validací příznaků nabízejí největší potenciál pro automatizovanou analýzu mikroskopických dat.

Klíčová slova

Strojové učení, Detekce objektů, Analýza mikroskopických obrazů, Farmakognozie, Vysvětlitelná umělá inteligence (XAI), Hluboké neuronové sítě, Augmentace dat, YOLOv10

Contents

1	Introduction	1
2	Related Work and Motivation	3
2.1	Structure of the Thesis	3
2.2	Algorithms for Computer Vision	3
2.3	Research Data Challenges	4
2.4	Machine Learning Applications	5
2.5	Deep Neural Networks in Microscopy	14
2.6	Summary and Significance	18
2.7	Contributions	20
2.8	Research Goals and Questions	21
2.9	Ethics and Reproducibility	22
3	State-of-the-Art	24
3.1	Foundational Concepts	24
3.2	Two-stage Faster R-CNN	32
3.3	Single-stage YOLO	33
3.4	Transformer-based DETR	34
3.5	Evaluation Protocols	35
3.6	Explainability	37
3.7	Practical Summary of Key Technologies	38
4	Methodology	39
4.1	Data Acquisition	40
4.2	Dataset Definition	42
4.3	Training Configurations	45
4.4	Baseline Models Configuration	46
4.5	Data Preprocessing and Augmentation	47
4.6	Novel Models Configuration	48
4.7	Training Optimization and Troubleshooting	50
4.8	Explainability Implementation and Configuration	51
5	Results	53
5.1	Quantitative Results	53
5.2	Qualitative Analysis: Explainable AI (XAI)	54
5.3	Summary of Key Findings	56
6	Discussion	58
6.1	Results Interpretation in the Context of Research Questions	58
6.2	Comparison with Prior Work	59
6.3	Practical Implications for Pharmacognosy	60
6.4	Limitations of the Study	61
6.5	Future Work	62
7	Conclusion	63

Bibliography	64
List of Abbreviations	77
Appendices	80

List of Figures

1	The increasing trend of published data papers from 2015 to June 2023. (Source: Giner-Miguel et al., 2025)	5
2	Comparative analysis of the AODSA-MPAVCP model with existing methods. (Source: Adam et al., 2025)	6
3	Comparison of mAP@0.5 outcome of AODSA-MPAVCP technique with existing models. (Source: Adam et al., 2025)	7
4	Improved vs. original YOLOv10 Loss Function. (Source: Zhao et al., 2025)	8
5	Results of gradient-weighted class activation mapping (Grad-CAM) ablation experiments. (Source: Zhao et al., 2025)	9
6	Comparison of sensitivity, specificity and F1-measure results among radiologist groups and also groups assisted by liver artificial intelligence diagnosis system (LiAIDS). (Source: Ying et al., 2024)	11
7	liver artificial intelligence diagnosis system (LiAIDS) algorithm architecture. (Source: Ying et al., 2024)	12
8	Organ-DEtection TRansformer (DETR) training architecture. (Ghahremani et al., 2025)	14
9	Detection results on the original dataset of microbial species by the R-CNN model. (Pawlowski et al., 2022)	15
10	Detected bacterial strains of <i>E. coli</i> and <i>S. aureus</i> by the custom you only look once (YOLO)v8s model. (Romphosri et al., 2024)	16
11	Predicted leukocytes in microscopic samples by the RF-DEtection TRansformer (DETR) model. (Dahiya et al., 2025)	17
12	Data analytics being used for a data-driven decision making. (Kelleher, 2020)	25
13	Main components of machine learning (ML): data, model and loss. Based on the continuous cycles of trial-and-error, each inference produces loss which is then used in subsequent runs, producing improvements of the model. (Jung, 2022)	26
14	Data being used to train a new predictive model. (Kelleher, 2020)	28
15	The geometrical representation of perceptrons with non-linear boundaries. From left: parabolic, circular and piecewise or step-like boundary. (Buduma, 2022)	29
16	Example structure of an artificial feed-forward neural network (FNN). (Kelleher, 2019)	30
17	Sample of data flow in LeNet Convolutional Neural Network. (Zhang, 2023)	31
18	The original R-CNN architecture diagram. (Girshick, 2015)	32
19	Comparison of NMS-based (LEFT) vs. NMS-free (RIGHT) architecture. (Chakrabarty, 2026)	33
20	Architecture of the DETECTION TRansformer (DETR). A standard CNN extracts feature maps, which are flattened and supplemented with positional encodings before passing into a transformer encoder-decoder sequence. The decoder utilizes a fixed number of learned object queries to directly output a set of bounding box predictions and class probabilities. (Carion et al, 2020)	34
21	The overarching methodology pipeline of the study, illustrating the end-to-end data flow from microscopic acquisition to statistical evaluation and Explainable AI validation.	39
22	Object count per image histogram	41

23	Label location heat map	42
24	Representative samples from the curated microscopic dataset, illustrating the high morphological variance and complex background textures typical of bright-field botanical preparations.	43
25	Ground-truth bounding box annotations applied to target samples. Precision in drawing the Regions of Interest (RoIs) is critical to prevent the models from learning spurious background artifacts.	44
26	Explainability techniques for the sample of <i>Althaea officinalis</i> : (a) center-only heatmap, (b) distributed mass over pixels heatmap, (c) cropped Grad-CAM for the highest predictions, (d) full image Grad-CAM visuals.	55

List of Tables

1	Reviewed models with reported mAP@0.5, source, and use case.	19
2	Reviewed microscopy models with highest achieved mean average precision (mAP), source and use-case.	20
3	Comparative summary of the evaluated object detection architectures and their practical characteristics.	38
4	Baseline object detection performance across evaluated architectures. . . .	47
5	Summary of generated dataset versions, highlighting preprocessing steps, augmentations, and resulting splits.	48
6	Summary of key architectural distinctions for the evaluated novel models. .	49
7	Per-class results on the full validation dataset, including 95% Wilson Confidence Intervals.	53
8	Per-class results on the consolidated dataset (v17) for the evaluated architectures.	54
9	Comparative analysis of the proposed YOLOv10x model against key prior studies in microscopic object detection.	59

1 Introduction

„We’re at a really exciting moment in history, I think, where we can use AI to build new tools for science... an AI-powered microscope or an AI-powered telescope... that allow you to see the world in a new way and see new truths about the universe.“ [1] — Demis Hassabis (Co-founder and CEO of Google DeepMind), in an interview with The Verge, 2021

The quality and power of tools used in science is what drives the future of research and development. Experiments with new materials, medicine and technology can be accelerated through the adoption of new approaches empowering contemporary solutions. The light microscopy has evolved allowing advanced techniques to capture the essence of studied materials in a time span of seconds, leading to new discoveries. Furthermore, where capturing and analyzing required a handful of qualified and talented individuals, today teams of multidisciplinary experts are needed to cooperate on evaluation of the multitude of data. As a result, the main effort has shifted from a small number of high quality images to a large-scale dataset which retains the same quality. Likewise the computational power has been expanding, creating opportunities to leverage its strength. Databases are able to maintain these datasets, allowing ease of access and a controlled environment for further processing and deep analysis. Consequently, the difficulty lies no longer in acquisition but in analysis of big-data. As a result, modern analytic tools need to provide qualitative and also quantitative capabilities. This is the most common bottleneck for collected scientific data today. Some of the latest solution proposals include a segmentation tool for multidimensional microscopy data or a complex platform for synaptic transmission in neurons, single-vesicle exocytosis and nano-sensor-detected exocytosis. [2, 3] Similarly to the focus of this thesis, both are fundamentally based on the state-of-the-art deep learning architectures.

The development of microscopic imaging technology can be followed from the 16th century, sparking the demand to advance its capabilities. Hence, the demand for exploration of the micro-nanoscience is responsible for the latest light, electron and ultimately the probe microscopes. [4, 5] More specifically, the light microscopy employed for the data collection in this work can be traced back to the first single lens microscopes, which further evolved into the applied compound microscope. Later technological landmarks in the light microscopy included creation of apochromatic lens for correction to the chromatic and spherical distortions and the first refractometer. Additional enhancement moved from the condenser unit to the Köhler Illumination which forms the contemporary standards of the bright-field light microscopy. [6] The 20th Century saw the production of medical microscopes with a series of rotary objectives, significantly advancing the speed with which the pathologists could change magnification. Finally, 21st Century microscopes can provide a water-,glycerol- or oil-immersion objectives allowing research of cellular details of a mouse embryo. [7] These are among few advancements which contribute to the growing number of progressively higher resolution multimodal imaging data, hence the increased requirement for analytical capacity. Failure to deliver advanced tools capable of supporting these needs, equals stagnation in research of drug discovery, diagnostics or material science. As a result, the global cross-disciplinary focus is on making such data findable accessible interoperable and reusable (FAIR), thus creating trainable collections of high quality annotated data for development of new generations of artificial intelligence (AI) algorithms. [8, 9]

Equally important is the computational power required to retrieve, process and analyze the microscopic data. Therefore, to address these challenges, the development of big data (BD) software is being actively spearheaded by high performance computing (HPC), machine learning (ML), deep learning (DL) and alternative AI approaches. [10] This would not

be possible without the preceding rapid technological growth from 8-bit microprocessors in 1970s to the modern 64-bit RISC architectures used today. Particularly, novel human-brain inspired architectures which are capable of outperforming the advanced tensor core H100 graphics processing units (GPUs) by magnitudes in floating point operations per second (FLOPS) and on-chip memory, both of which have the potential to significantly advance the analytical operations. [11] However, the Moore’s Law has been slowing. Furthermore, the year 2018 detected an approximately 15-fold divergence from the Moore’s original forecast. [12] Since, the physical limitations of the contemporary hardware (HW) are being estimated, it is critical to elaborate and leverage the strengths of current AI systems. Moreover, the specialized HW has evolved from the pure GPUs, through tensor processing units (TPUs) to neural processing units (NPUs). Consequently, creating the drive for innovation in future HW and software (SW) architecture co-designs to include the latest ML techniques, e.g., the deep neural network (DNN)s. [13, 14, 15] Such systems depend on complex analytical pipelines. These include data preparation, training and inference to produce valuable statistical insight. Resource consumption and the time required for execution are critical parameters for benchmarking in state-of-the-art SW and HW. [16] In addition, with the constant evolution in these fields, multimodal AI is enveloping areas of text, visual and audio variables. Such improvements allow for remote expert analysis in regions without specialists in fields such as computational diagnosis and biological pathology [17, 18].

To summarize, impactful results require powerful tools and interdisciplinary participation. Recent advancements are primarily focusing on application of various DNN algorithms, like the convolutional neural networks (CNNs) or the recurrent neural networks (RNNs), which learn and detect objects with increasing efficiency. [19] Although this thesis focuses on microscopic object classification and detection, the DNN architectures are applied in a broad spectrum of operations. These include aerospace defense and cybersecurity for detecting modern unmanned aerial vehicles (UAVs) and for development of autonomous maritime vehicles. [20, 21] Furthermore, financial sector can capitalize on price forecasting, algorithmic trading, risk management, credit scoring, customer segmentation, contract analysis and, with certain degree of success, stock market predictions. [22, 23, 24] More applications are discussed in later section on related work and research context. Nevertheless, the examples mentioned up to this point demonstrate that data, informatics and computer science and technology are interdependent and mutually reinforcing with any other hosting discipline. Thus, cutting-edge projects require the collaboration of several specializations to overcome the complex challenges of each. These might include imaging experts operating the sensory devices, biologists and biochemists providing the ground-truth about the retrieved samples, data scientists building the detection models, etc. Such collaborations are the optimal way to successfully conquer the large data bottleneck described earlier. As a result the extension of analytical capacity grants the access to new and refined *truths about the universe*.

2 Related Work and Motivation

The main focus of this work is to develop, train and evaluate different ML models which are applied for complex microscopic object detection. Such approach opens up ways to expedite analysis of understanding the underlying patterns in brand new data which might have otherwise remained obscured. Constructed architectures are based on the selection from distinct approaches to the most often utilized DNNs for analysis in the fields of medical and microbiological image data, as suggested in the introductory part. The secondary ambition of this thesis is to experiment with the applied architecture designs to achieve optimized results for the studied problem in microscopic object detection and classification. Generally, this proposes optimizations for small and medium targets, to correctly localize and classify each. Additionally, the added benefit of creating the open access dataset with labeled and validated bio-image data is in line with the latest global scientific initiatives for re-usability of shared research data. Finally, the constructed models provide a base for future research into the area or can be integrated into another multi-modal system solution.

2.1 Structure of the Thesis

To systematically address the defined research goals and introduce the novel, FAIR-compliant bio-image dataset, the remainder of this thesis is logically structured into six subsequent chapters. Chapter 2 establishes the theoretical foundation, reviewing contemporary literature on object detection architectures and the current state of deep learning within microscopic image analysis. Chapter 3 details the rigorous methodology behind data acquisition, annotation, and the iterative pre-processing strategies (versions v6 through v17) deployed to mitigate inherent class imbalances. Chapter 4 outlines the specific architectural configurations and training protocols for the YOLOv10x, Faster R-CNN, and DETR paradigms. Chapter 5 presents the empirical results, utilizing strict quantitative metrics and qualitative Explainable AI (XAI) techniques to transparently validate the models' feature-extraction capabilities. Chapter 6 synthesizes these findings, discussing the practical implications of utilizing highly optimized, single-stage detectors for hierarchical botanical sub-classification, while also addressing study limitations and proposing future avenues of research. Finally, Chapter 7 concludes the thesis by summarizing the core empirical victories and their broader significance to the automation of pharmaceutical botany.

2.2 Algorithms for Computer Vision

This work explores the (i) you only look once (YOLO), (ii) faster region based convolutional neural network (R-CNN) and (iii) DEtection TRansformer (DETR) algorithms. YOLOv10 employs a large-kernel convolutions for a single-stage object detector. Key innovation in the version 10 relies on removal of the need for non-maximum suppression (NMS) during inference. As a result, the throughput is optimized for FLOPs and sensitivity to hyper-parameters. This modernization arose from the challenge posed by the DETR designs. [25, 26]. The faster R-CNN operates on pre-generated region proposal network (RPN) before activating classification and bounding box regression. The RPN serves as an attention network for identifying the regions which should then be processed. Initially, the input is refined by the RPN multi-convolutional backbone before it is transferred to the detector. There the image is classified based on the proposed regions. For time optimization, the RPN and the detector share these convolutional layers [27, 28]. Lastly, the DETR architecture is an end-to-end, single stage detector based on self-attention mechanisms and

distinctive encoder-decoder structure. The original success of the transformers in the field of natural language processing (NLP) has encouraged further exploration into image classification applications. DETR is the pinnacle of the original Transformer architecture due to overall computation and potent storage capabilities. Moreover, bypassing the traditional anchor boxes or NMS produces powerful streamlined process. As a result, utilizing attention mechanism directly detects objects, confirming the locations in the processed images. In addition, DETR generally allows for employment of greater number of parameters than CNNs which grants it more versatility during detection tasks. Ultimately the main advantage lies in its sparse detection, which is in contrast with the established dense detection. This is achieved through transformation of the input problem into an un-ordered sequence of outputs [29, 30, 31].

Each of these modern algorithms provides different sets of advantages and disadvantages, hence understanding how each operates is crucial for the right deployment. Furthermore, to design any specific optimization or improvements, intelligence of the inner workings is required. Thus, a detailed study is provided in the subsequent state-of-the-art section.

2.3 Research Data Challenges

The domains of nature and medicine are the primary source of microscopic data collected to create foundation models in these fields. However, data collections available online are still suffering from insufficiency in training data for the use in experimental data-hungry ML models [32]. Open science initiatives are becoming increasingly welcomed by the scientific community due to the European Union (EU) open access directive 2019/1024 to the research data collected [33]. As a result, terabyte-degree collections are being funded and accepted, containing multidimensional bio-imaging data from electron and light microscopy [34, 35].

During the examination of 4041 research articles containing data used for ML in various fields over the time span of 8 years, a growing trend, seen in Figure 1, of data availability is evident. Nevertheless, out of the 10 most prominent topics only 330 were medicine, 202 material, 160 biological, 87 agricultural and 72 drug related, amounting to 8.1%, 4.9%, 3.9%, 2.1% and 1.8%, respectively [36]. To surpass the shortage of data, unsupervised techniques can be employed. However, previously tested weakly supervised ML approaches produced low accuracy results, mainly due to the difficulty of defining the region of interest (ROI) and overlapping targets creating mixed results [37, 38]. The key challenges the DNNs are facing in this area are the lack of comprehensive and annotated training datasets and the absence of pre-trained model weights for transfer learning options. Additional difficulty is created by the diverging data collection techniques, hence creating problems for the models to successfully generalize on new data. This can be solved using various views, sizes, modalities, light and color conditions, magnification and further data augmentations [39, 40].

Once the most time-consuming tasks of collecting and annotating data are completed, another impediment is the lack of experts willing to validate the annotated datasets, creating inconsistencies. [41] Furthermore, the processing power requirements for the advanced ML are still expensive relying on evolving HPC resources, suggesting discouragement for future investments. [42] Finally, one of the greatest complications the modern DNNs present is the interpretability and reliability. The obstacle has evolved to its own field denoted as explainable artificial intelligence (XAI). The critical source of trust issues is the inherent black-box nature of these architectures. Particularly important is this issue in clinical practice where understanding the reasoning behind diagnosis or prescription is of paramount

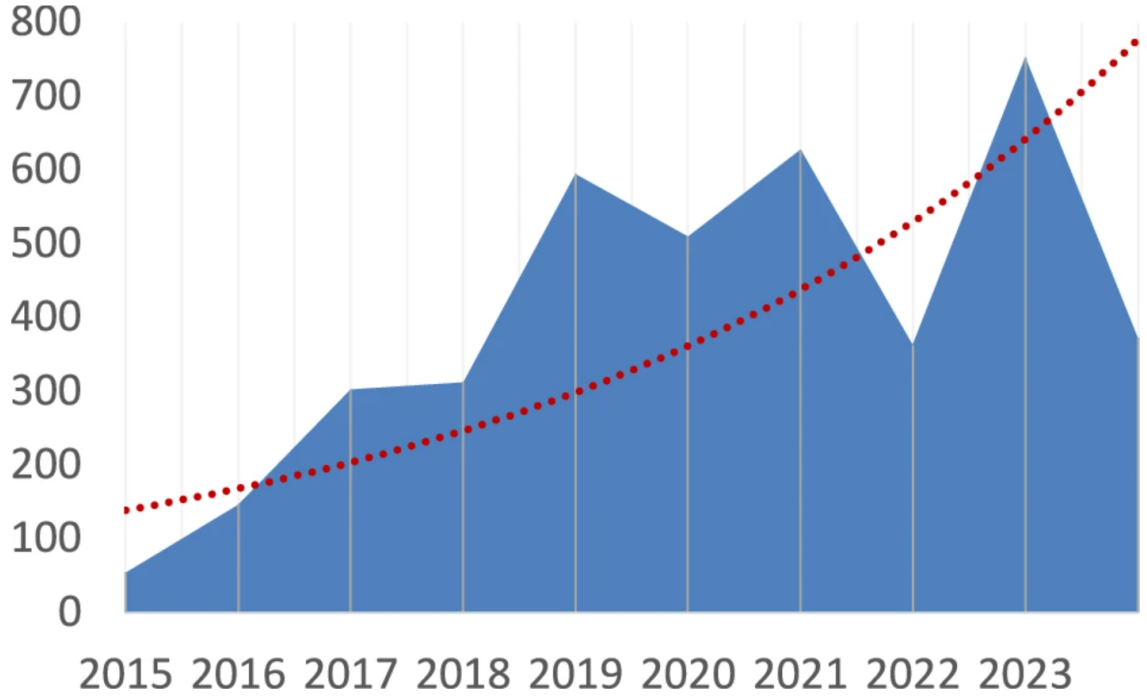


Figure 1: The increasing trend of published data papers from 2015 to June 2023. (Source: Giner-Miguel et al., 2025)

concern. [43] The latest solutions revolve mainly around the ability to display confidence levels or intervals, with the aim to develop novel feature visualization capabilities. [44, 45] Currently, the state-of-the-art techniques employed in XAI are based on gradient-weighted class activation mapping (Grad-CAM). Nevertheless, the question whether the highlighted regions within the heatmap outputs of the Grad-CAM correspond to the relevant clinical features remains [46, 47].

2.4 Machine Learning Applications

One key area where computer vision (CV) assistance is able to provide immediate benefit is aiding the visually impaired persons. Rapid detection and classification of objects is a daily task general populace performs every second. However, fading or lost sense of sight can be life threatening in areas such as navigating the daily traffic, encountering new environments or changes in the default routes. This is an example which requires a real-time problem solving. In a recent study on advanced object detection for smart accessibility, the research team focused on utilizing the YOLOv10 algorithm for the purpose of developing a new model specifically focused on aiding the visually challenged. The study proposed the advanced object detection for smart accessibility using the marine predator algorithm to aid visually challenged people (AODSA-MPAVCP) [48]. The model produced potent results in $prec_n$, $reca_l$, $accu_y$, and $F_{measure}$ of 95.31%, 89.96%, 99.63%, and 92.15%, respectively as seen in Figure 2. For the mean average precision (mAP) the study uses the metric of 50% intersection over union (IoU) threshold, which equates the overlap between predicted and ground-truth boxes denoted as mAP@0.5. In this metric the model AODSA-MPAVCP

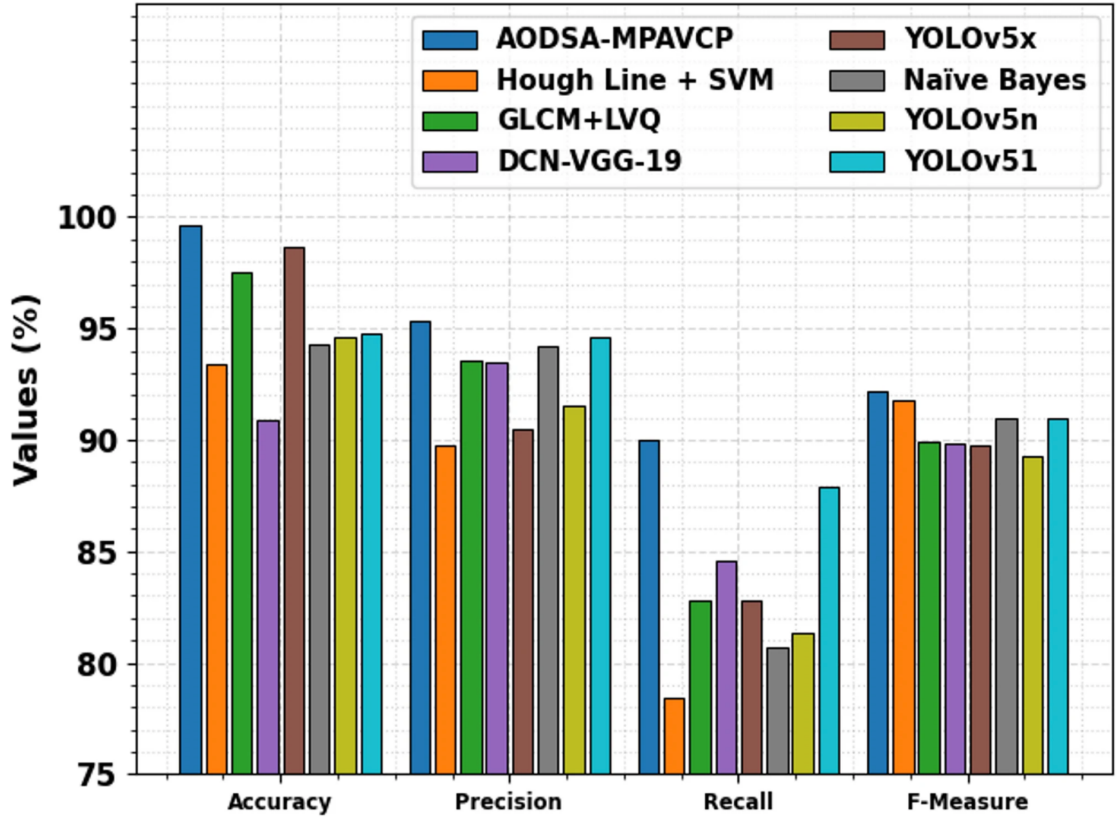


Figure 2: Comparative analysis of the AODSA-MPAVCP model with existing methods. (Source: Adam et al., 2025)

managed to reach 78.83% as shown in Figure 3, outperforming all other specific models by a significant margin. It is worth noting that mAP@0.5 metric is overall generally accepted as being the most accurate for the detection results, despite its lack of negative impact generated by the false positive (FP) findings. Consequently, the F1-score tends to be more effective in task specific fields of detecting and classifying image data [49, 50].

As their proposed object detection model, the team utilized the YOLOv10 architecture with feature extraction process being powered by the deep convolutional neural network by Visual Geometry Group at University of Oxford (VGG19). This converts original input data into relevant features. Furthermore, the deep belief network (DBN) method was applied for the classification task and to optimize the outputs, the marine predator algorithm (MPA)-based hyperparameter tuning was executed [48]. The MPA is inspired by the feeding patterns of marine organisms and was developed as an intelligent optimization algorithm which benefits from a straightforward structure and a few parameters, hence its accessibility. On the other hand, the MPA suffers from untimely convergence and inadequate diversity [51, 52]. The DBN is in essence a multi-layered feature extractor which is capable of determining patterns inherent in the data. Power of this network originates from the end of each layer where the system learns the entire structure from the given input. Comprising hidden units, denoted as latent variables, these units are connected between the layers and not within individual layer. This is in contrast with general DNNs where this process

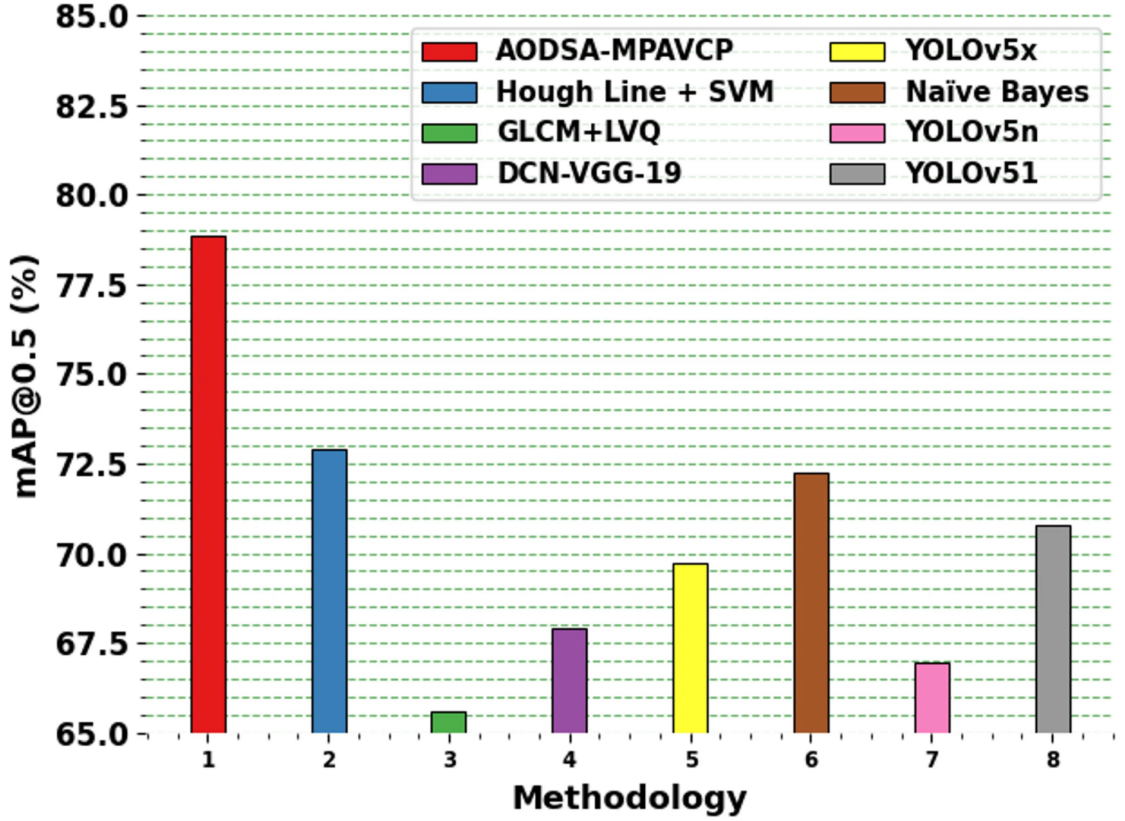


Figure 3: Comparison of mAP@0.5 outcome of AODSA-MPAVCP technique with existing models. (Source: Adam et al., 2025)

happens progressively with each stage in the network, a DBN solves the gradient blur issue by learning the hidden features and patterns at the general level. Unsupervised phase of DBN is often followed by supervised training for applications in detection and classification. [53, 54] Finally, the combination of YOLOv10, VGG19, MPA and DBN was successful in reaching the overall accuracy of 99.63% for the indoor object detection dataset by the AODSA-MPAVCP model [48].

Second major area where real-time object detection carries significant value is defense and security. Whether it is used in autonomous vehicles to navigate against road hazards or to warden natural resources from environmental risks. Furthermore, face recognition plays a valuable role during surveillance, aiding in identification of missing persons or monitoring unauthorized entry. Security cameras are being installed in places advantageous to deter potential conflicts and provide tangible evidence in case escalations arise. It is worth mentioning that despite its added benefits, the increased technological surveillance has also been linked to increased psychological pressure and decreased mental well-being in certain individuals [55, 56]. However, due to an increasing rate of safety incidents of various degrees and a steady decline of psychological health at schools or campuses [57, 58, 59] another research team has focused on a smart campus danger detection system. In the study, the team integrated a self-calibrated illumination algorithm to boost detection accuracy in low light conditions. The model is based on YOLOv10 network augmented by StarNet to decrease the number of parameters, calculations and to enhance the feature extractions.

Improving the YOLOv10 algorithm has lead to a decrease of the classification loss of the new model by approximately 20% and the feature point loss by nearly 16%. The experimental results can be viewed in Figure 4, displaying the stabilized loss values below 0.6, 0.4 and 1 for boundary frame loss, classification loss and feature point loss, respectively.

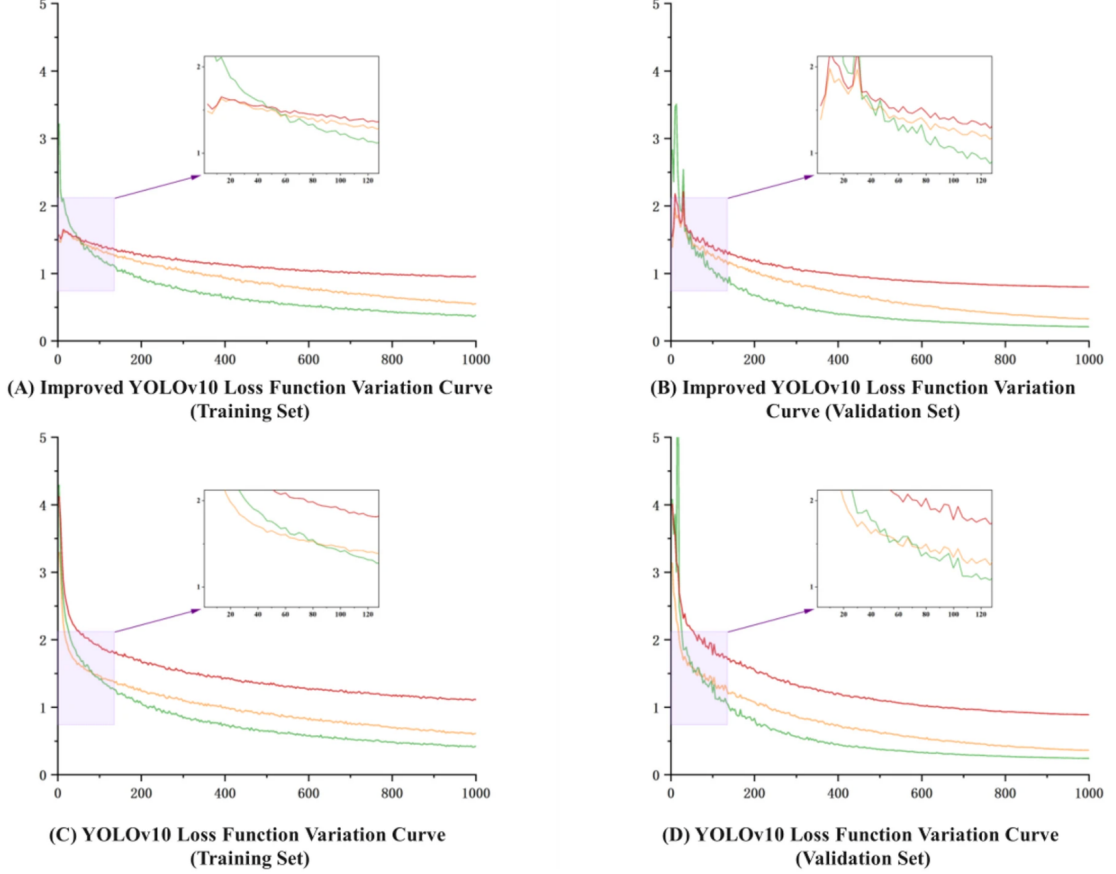


Figure 4: Improved vs. original YOLOv10 Loss Functon. (Source: Zhao et al., 2025)

Loss functions are critical for calculations of the difference between the ground-truth and prediction results. Loss functions can be generalized as an objective functions outputting the scalar measure of the machine learning model efficacy [60]. Moreover, to profoundly inspect the impact of each specific component added to the original algorithm, the team performed ablation experiments. In regards to the XAI, the study included a Grad-CAM visualization model. Since the Grad-CAM localization procedure is gradient based it is capable of displaying the relative importance by an image-classification CNN model [61]. The heat-map ablation experiment can be seen in Figure 5.

This approach illuminates the prominent areas of attention to assist with explainability of the new method. They also performed the ablation experiments for auxiliary bounding box, StarNet and convolutional block attention module (CBAM), as the proposed enhancements. StarNet is a novel model which supports multi-scale spatio-temporal feature extractions utilizing the CNNs. Which allows to learn both images and video sequences of patterns at multiple levels of detail. [62, 63]. Similarly, the CBAM incorporation aims to improve the feature extraction by spatial and channel-wise pattern detection. The attention

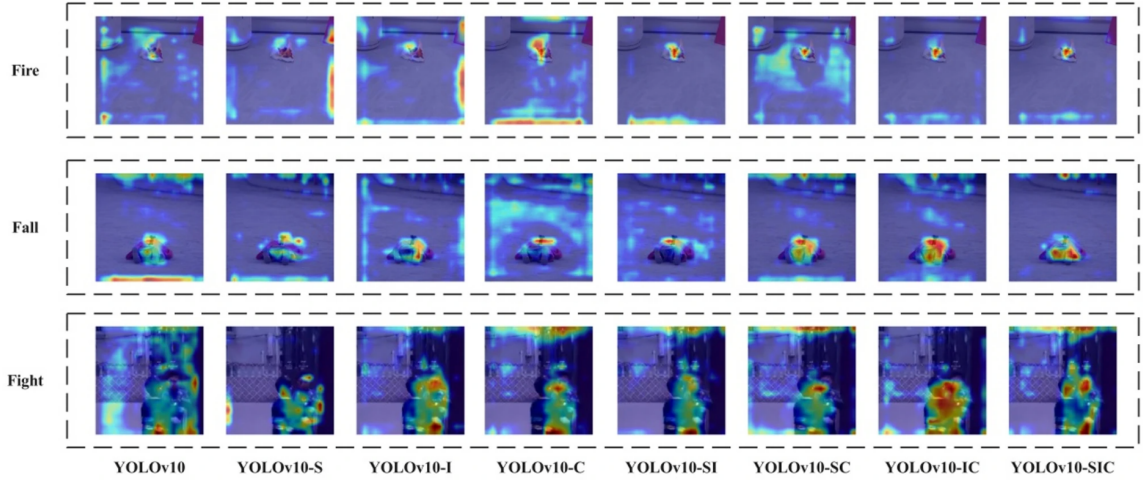


Figure 5: Results of Grad-CAM ablation experiments. (Source: Zhao et al., 2025)

mechanism inside the CBAM amplifies crucial aspects. Hence, both of these additions are designed to specifically advance the object detection capabilities in a video sequence [64, 65]. In summary, this has lead to the number of model parameters drastically decreasing and accuracy slightly increasing in comparison with the original YOLOv10 due to the enhancements. More specifically, the parameters were reduced by over 80% meanwhile increasing the precision, recall, F1-score and mAP by 0.99%, 3.31%, 2.15% and 1.23%, respectively. [66]

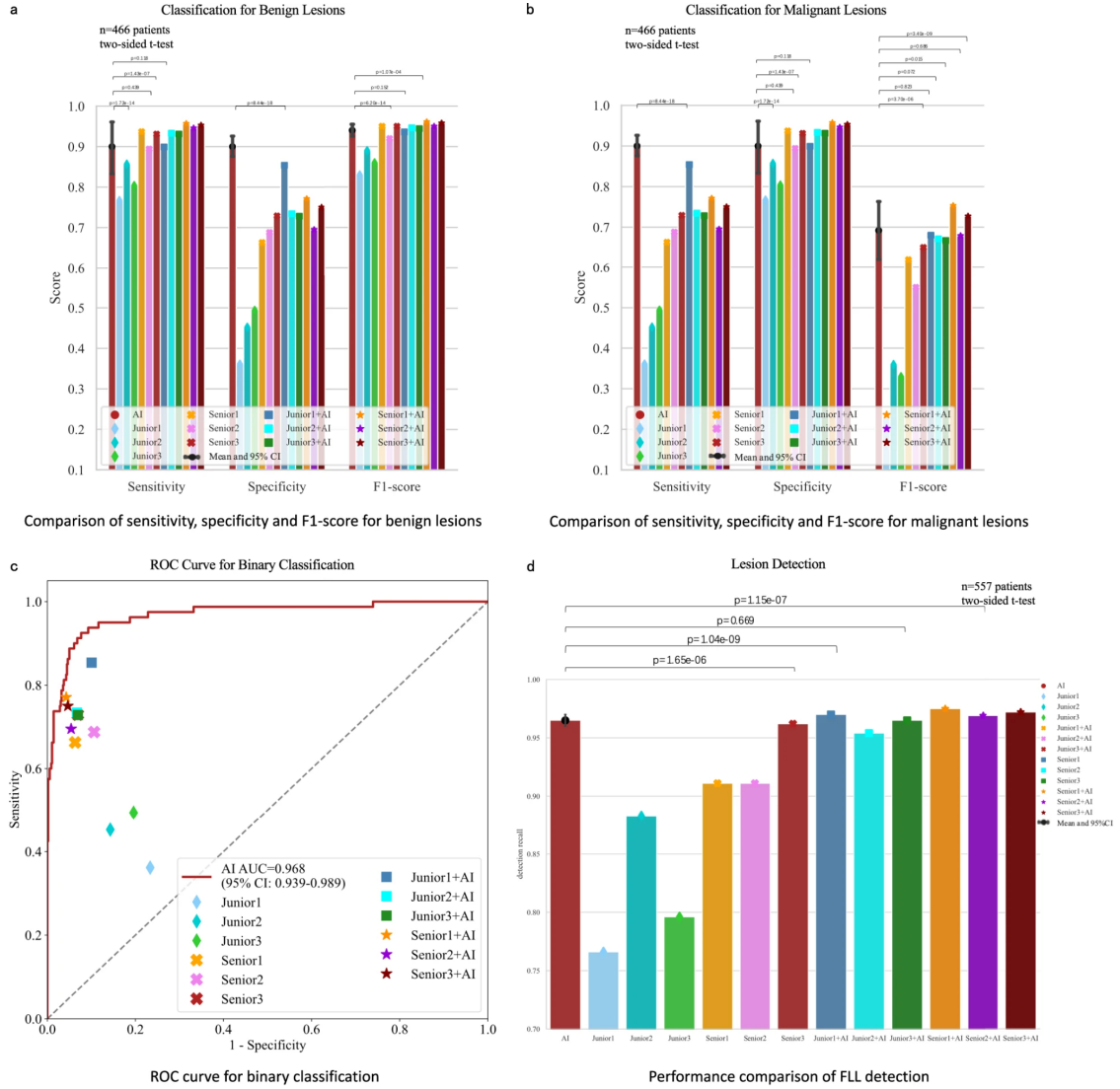
Previously mentioned applications were focused on the real-time detection. Nonetheless, the employed YOLO algorithm with its myriad of versions has already proven the capacity to outperform two-stage detectors like Faster R-CNN or DETR in accuracy [67, 68, 69, 70]. Hence, implicating broader compatibility with future improvements and custom builds. For instance, this has been recently tested in driver assistance systems for use in foggy environments, here the custom made HR-YOLO method reached 79.8% in mAP@0.5, significantly outperforming Faster R-CNN at 66.9% and DETR at 67.5% [67]. Another example performed the experimental analysis for non-invasive neuroblastoma diagnostics approach, where the team applied 4 variants of YOLO model. These were: original v8, and then 3 custom builds v8-iRMB, v8-EVC and v8-IE, with final results reaching mAP@0.5 of 81%, 82%, 85% and 89%, each. The main enhancements are attributed to the centered feature pyramid network (FPN) structure. This results in improved feature extraction for objects with varying size and shape. Furthermore, when compared to the Faster R-CNN the YOLOv8-IE surpassed it by 7% in mAP@0.5 albeit the number of parameters was decreased by 156.96 millions, significantly shrinking the size of the model [68].

In spite of these findings, the larger algorithms like the Faster R-CNN can still provide more reliable and robust information for specific non-temporal tasks. Furthermore, this is essential for scenarios where the accuracy and precision outweighs the need for velocity. There, the emphasis is on gathering the most relevant aspects of the investigated material or underlying patterns with little regard to analysis duration. Such areas include inspections and diagnosis, for instance the remote sensing satellite images, defect disclosures or endangered species detections [27, 71, 72]. In the industrial steel surface defect detection, the team enhanced the baseline Faster R-CNN model by integrating feature fusion module

between RPN and FPN. Resulting in improvement of mAP by 1.6% by boosting the feature extraction capacity in fine-grain detailed images. Additionally, by adding lightweight channel attention mechanism and also replacing the standard convolutions with the depth-wise separable convolutions, the parameters of the model were reduced the model complexity. Consequently, decreasing the size by 1419KB, accelerating the inference speed by 3.8% and further pushing the final mAP to 80.2% [71]. In remote sensing from satellite images, the research was primarily conducted to develop bi-stage compression approach to retain the performance of the Faster R-CNN while allowing the decrease computational and storage requirements. The key modifications that the team applied to achieve this were adopting the mixed-precision FP16 computation model for training and then utilizing the unstructured weight pruning along with dynamic quantization in the post-training phase. Due to these changes, the definitive R-CNN model with the mAP of 95.4% was able to outperform all competing YOLO8n, YOLO8m, YOLO10m, and YOLO11m systems which attained 68.6%, 71.02%, 72.5%, and 67.37%, respectively on these datasets. However, the training times diverged drastically with 1.8 h for the R-CNN and 0.283, 0.464, 1.128, and 1.142 h per each YOLO model [27]. As a result of these experiments, the two stage detectors retain the relevance in specific fields. This is especially true for crucial fields like medicine, where early diagnostics of breast tumor or an autism spectrum disorder have the power to radically impact human lives [73, 74].

Such medical approach was adopted by a recent study for early detection and diagnosis of focal liver lesions. Responsible research team has developed and validated AI solution denoted as the liver artificial intelligence diagnosis system (LiAIDS). The final model achieved the F1-measure of 94% for benign lesions and 69.2% for malignant cases [75]. Additionally, the system was contrasted with F1-measures of designated groups of junior and senior radiologists in this field. Furthermore, these professionals had then been assisted by the LiAIDS model and conducted new set of assisted diagnosis, the results can be seen in Figure 6. The imagery data, collected by 18 hospitals comprising 11,385 patients, were applied as the initial training data for development and validation of the LiAIDS system. Later, another 1,225 participating patients were part of the prospective validation support group. For the training data preparation, only experienced general radiologists who obtained at least 5 year-experience in the focal liver lesions diagnosis were requested for the image annotations and divided into 5 paired groups. The ground truth standard was determined based on historical histopathological reports or confirmed by senior radiologists with more than 20 year-experience in the field. In each group, one radiologist would be responsible for boundary definition while the other played the supervising role.

Three main modules, the lesion detection, liver segmentation and lesion classification, compose the main architecture of the LiAIDS model. As displayed in Figure 7, localizing and determining possible candidate regions in computed tomography (CT) scans was performed by the lesion detection module. This sector is powered by the Faster R-CNN framework with augmented cross-shaped window (CSWin) Transformer serving as the feature extraction backbone. Similarly to previously discussed studies, the team integrated the FPN into the RPN to boost spatial detector performance. Then false-positive filter was provided by the segmentation module, removing non-relevant lesions for the liver area. The encoder-decoder U-Net architecture along with scSE attention and Non-Local attention mechanisms were adopted to enhance segmentation accuracy. The last module was designed to separate detected classes into malignant or benign based on the extracted features compared with the clinical information. The composition of DenseNet and scSE attention unit create the feature extraction backbone. The outcome of the study reports that the LiAIDS



AI=Artificial Intelligence; 95%CI = 95% Confidence Interval; ROC=Receiver Operating Characteristic; AUC=Area Under Curve; FLL=Focal Liver Lesions.

Figure 6: Comparison of sensitivity, specificity and F1-measure results among radiologist groups and also groups assisted by LiAIDS. (Source: Ying et al., 2024)

model outperformed the junior and attained comparable results with the senior group of radiologists. For this experiment, six differently experienced general radiologists were divided into two groups based on their levels of expertise. These were junior with 5-10 and senior with 10-20 years of experience. In the contest against the system solution, 557 randomly selected patients were chosen. The trials were run for each group separate and then also in cooperation with the model. Consequently the F1-measures of the junior group were 0.830, 0.890, and 0.860 for benign lesions and 0.230, 0.360, and 0.330 for malignant instances, this is on average worse than the LiAIDS model by 0.080 for benign and 0.385 for malignant samples. Results of the senior group was on average identical with the model for benign and only by 0.085 worse for the malignant cases. However, when paired with the system for a collaborative study, the F1-measures drastically increased for the junior group and

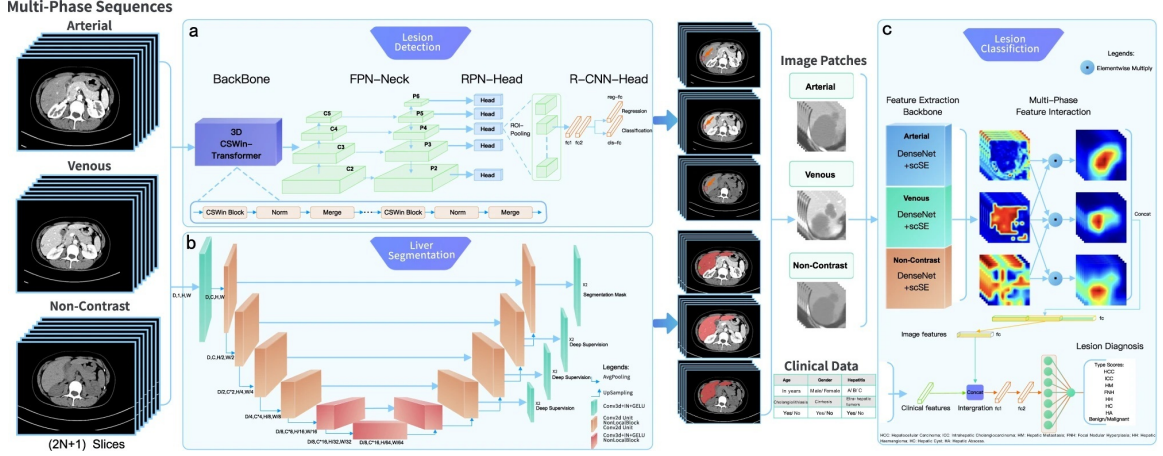


Figure 7: LiAIDS algorithm architecture. (Source: Ying et al., 2024)

the senior group managed to attain a score superior to all previous testing. Collaboration of the senior group with the AI produced F1-measures of 0.961, 0.950, and 0.958 for benign and 0.753, 0.679, and 0.727 for malignant ones. To summarize, this advanced automated model can be deployed as a powerful tool bolstering the human accuracy in detection and classification of liver lesions. Additionally the LiAIDS system can be utilized during training and education of new radiologists or serve as an invaluable consultancy tool in regions with scarce number of these professionals [75]. Ultimately, further enhancing the quality of life and improving health expectancy of all humans.

The feature extraction backbone CSWin, employed in the previous study, belongs to the family of vision transformer (ViT) architectures. This specific build relies on self-attention mechanism which processes input features into stripes of equal sizes. Specifically, computes in parallel the self-attention in horizontal and vertical stripes, forming a cross-shaped window. Furthermore, the CSWin transformer operates with locally-enhanced positional encoding (LePE) enforcing the attention result to be processed based on the positional information in each transformer block. Hence, making it a potent tool in object detection tasks. [76] Due to the marvelous achievements in the field of NLP the Transformer technology is actively being explored for further application opportunities, especially for the CV research. Allowing the modeling of long input dependencies while maintaining the paralleled sequence computing is one of the core assets, directly opposing RNNs, e.g., Long short-term memory architectures. What separates the transformers from CNNs is the minimum number of assumptions about the world, e.g., that nearby pixels relate to one another. Therefore, the relationships between pixels are derived via the attention mechanism, instead of pre-imposed locality. Since Transformers accept the inputs as an unordered set of tokens, similarly to NLP tasks, they remain flexible in global relations. Such potent architecture where every input token attends to every other token without forcing the model to focus on neighboring pixels is capable of multi-modal processing, i.e., text, speech, images or videos. Ultimately resulting in the first object detection application of the ViT architecture as a backbone to create the DETR model [77, 78, 79, 80].

Practical applications can be seen in the late medical study on detecting early signs of diabetic retinopathy. As a prominent origin of vision loss in diabetic patients, the earliest symptoms are as important as they are difficult to uncover. For the reasons described

earlier, the research team adopted the real-time detection transformer (RT-DETR) architecture which further excludes the NMS and adds multi-scale extractor to improve tiny and dense feature detections including varying scale targeting. For the experimental analysis, a comparative survey with four YOLO models was performed. Resulting in 88% score in mAP@0.5 metric outperforming YOLOv5 by 4%, YOLOv8 by 2%, the standardized DETR by 5% and single-shot detector (SSD) by 10% [81]. Thus, confirming the transformer-based models potential contribution in automated medical diagnosis. Furthermore, another research team has concentrated on differentiating the performance between medical and natural image data. Some of the key concerns were the variability in image resolution, making targets harder to detect, scarce ROI and the ambiguity among similar classes. Additionally, demonstrating the contemporary techniques for accuracy improvements can prove detrimental when dealing with medical and microscopic data as opposed to natural images. As a baseline model the team used Deformable DETR with Swin-T backbone in its default setting [82] and the AdamW optimizer [83] to reduce the learning rate by a factor of 0.1 during the remaining 20 epochs of training. Observations the team has made suggest these models tend to perform well on localization but poorly on classification tasks. The deformable DETR achieved mean average precision across IoU thresholds from 0.10 to 0.50 in steps of 0.05 highest of 49% with image resolution scaled to 0.75. During testing of different encoder layers the top result was 67.7% under the Swin-T backbone with 3 layers of encoders, outperforming all variants with ResNet50 backbone. Feature levels ablation assessment unveiled that DETR with all 4 default feature maps performed the best, also arriving at 67.7%. In summary the team recommends focusing on other DETR architecture designs capable of processing full-resolution images and handling overfitting during training caused by limited positive indicators [84].

Such trend points toward further investments into simpler light-weight solutions and classification modules instead of constructions of overly complex and robust solutions. In fact, the current course of action confirms this by heavy focus on portable and adaptable applications. For instance, the MobileLiteNet deployable on smartphone devices for traffic safety, the lightweight-YOLOv8 with lightweight-HGNetV2 backbone for efficient use in urban environments with limited computational resources or the detection algorithm Focus-Det, which aims at wider range and lower level features by discarding redundant prediction boxes in the post-processing stage. [27, 70, 85, 86] Thus, the ML models must be able to perceive the patterns in any provided scale and also offer invariance to translations across all layers, not only in the selected few.

Following the previous article, a recent study aimed at construction of a 3D organ detector, its framework is visible in Figure 8. Similarly, they based their model on the DETR transformer architecture. However, to combat the limitations of their predecessors, two new components multiscale attention (MSA) and dense query matching (DQM) were integrated. The new Organ-DETR model first inputs the CT image, a set of feature layers is provided by the residual neural network (ResNet) and FPN convolutional neural network (ConvNet) backbones. Out of this collection of layers, the highest one is selected and inputted into MSA. Inside this encoder, an adaptive spatial sampling occurs, i.e., patterns are reiterated through different scales. Two main mechanisms for this process are multi-head self-attention for intra scale attention map and multi-head cross-scale attention which examines embeddings on different scales. Deformable DETR decoder then filter positional queries and the last layer outputs the bounding box coordinates. Finally, the DQM is utilized for matching ground-truth queries to those predicted by MSA. Multiple medical datasets comprised the competing grounds for multiple methods with different backbones, transformers and mather

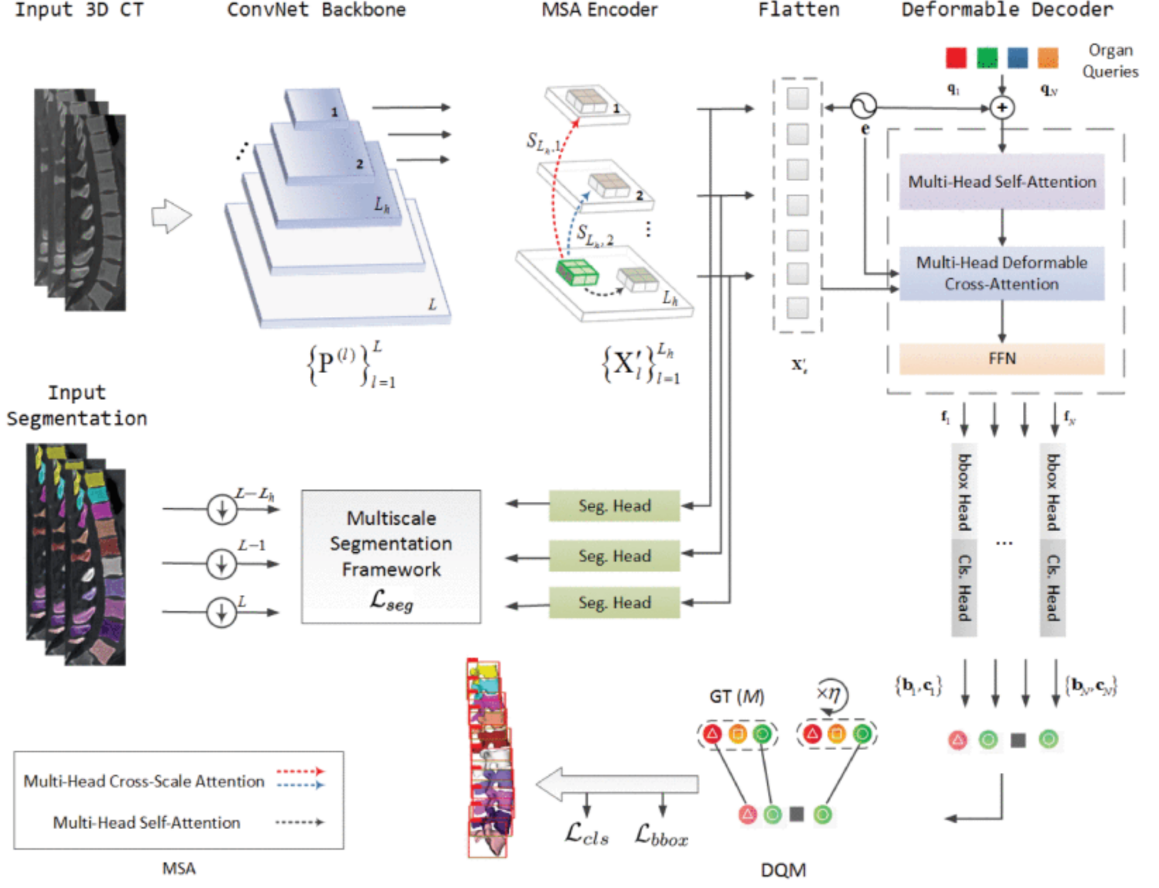


Figure 8: Organ-DETR training architecture. (Ghahremani et al., 2025)

modules. The new Organ-DETR managed to triumph among them all. When averaging the results over all testing datasets the new architecture resulted in mAP, mean average recall (mAR) and mAP@0.75 with 54.3%, 61.4% and 56%, respectively. In fact, compared to the best outcomes from all other contesting methods across all image datasets, the results were better by 10.6%, 10.2% and 13.8% per each metric. As a result, this structure provides a robust and effective performance across variable data, scale and specifications, which is in line with the recommended approaches from the previously studied article [84]. The authors of the Organ-DETR summarize by suggesting that a U-Net-Based ConvNet backbone, e.g., the FPN, should be capable to outperform ResNet, despite its considerable differences in size of parameters and computational time [87]. Thus, reinforcing the state-of-the-art paradigm of light-weight solution development, discussed earlier.

2.5 Deep Neural Networks in Microscopy

Based on the previous review of related work and due to the targeted application being object detection in microscopic imagery, the following architecture families were selected for this work: Faster R-CNN, YOLO and DETR. Despite the power of the latest models, the ability to generalize and to transfer learning from one problem domain to another can be limited. This is especially true in fields of microbiology, where sample tissues can differ

exponentially. Furthermore, the annotation work done by trained professionals is expensive and high-quality material is scarce. Captured image data can also face unexpected artifacts or scaling challenges resulting in unreliable predictions and biases. The ROI and targeted objects can overlap with artifacts, accumulate and deform each other, creating a difficult environment [88, 89]. All of these potential complications warrant a broader ML inspection to assess the power and weakness of each solution for the given task. The goal is to provide a functioning prototype and a detailed analysis for each, allowing the possible future multi-modal integration into real world projects.

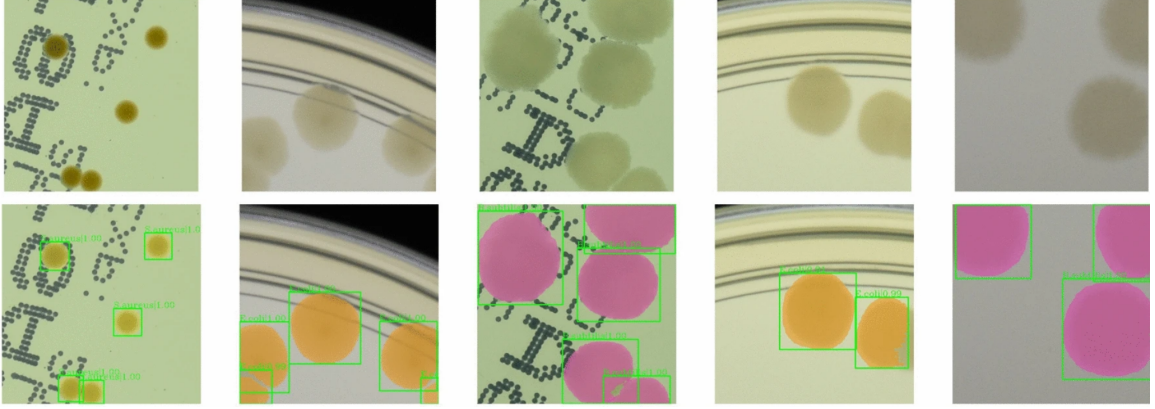


Figure 9: Detection results on the original dataset of microbial species by the R-CNN model. (Pawłowski et al., 2022)

An example of Faster R-CNN application in microscopy is the study to develop synthetic microbiology dataset generator and compare the DNN trained on the synthetic vs real data. The example results on real samples with different microbial species are displayed in Figure 9. First, the team developed the generator which is based on CV methods coupled with data augmenting neural transfer method. Generated synthetic dataset was then used for training two Faster R-CNN models and their results contrasted with the outcomes achieved on purely authentic data. Synthesizing 50k microbial patches required only 100 original images to generate. Finally, the models trained on the manufactured data reported only slight decrease of performance than those trained on the original 7k real images, which constitutes 65k of microbial patches. The two models were Faster R-CNN and Cascade R-CNN, measured in $mAP@0.5:0.95$, with the averaged detections of 0.401 and 0.416 for the synthesized and 0.493 and 0.520 for the original datasets, respectively [90]. The team has effectively demonstrated the plausibility of training DNN on augmented data to reduce the human resource strain on acquisition and labeling. Even though the results were slightly worse than training on the true data, it creates more options for constructing a more robust and flexible end-product. Additional 10k patches to the already fabricated dataset did not improve the detection fidelity, suggesting limitations to this method. Nonetheless, this technique stands as a compelling strategy applicable to detection and classification in microbiology.

A recent application of YOLO algorithm, specifically in microbiology, was during development of a rapid bacterial identification system. The research team focused on creating a ML solution which could expedite the medication of sepsis-like infections. Consequently, the speed was the primary motivator which led to selecting YOLOv8 as the ideal architec-

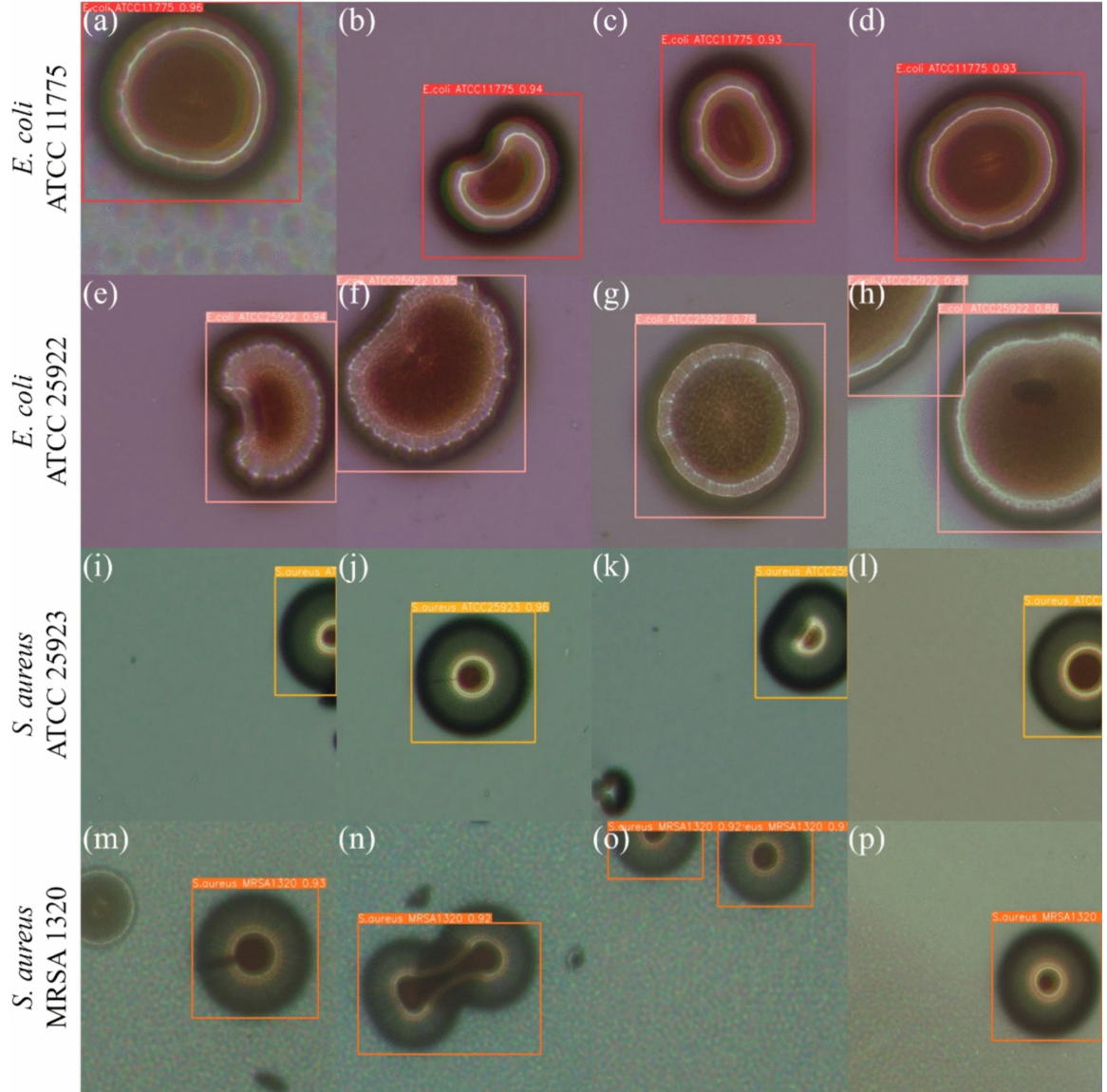


Figure 10: Detected bacterial strains of *E. coli* and *S. aureus* by the custom YOLOv8s model. (Romphosri et al., 2024)

ture, similarly to the reviews in previous chapters. As the primary model backbone the cross stage partial (CSP) design was chosen. For feature extraction the YOLOv8 network integrates the spatial pyramid pooling-fast (SPPF) module to increase the inference speed and the C2f module, which adds an extra skip-layer connections and split operations. Thus, allowing the model to retain the lightweight construction while capturing the enhanced flow of gradient [91]. For the neck of the YOLOv8, the team used the standard FPN as explored previously and the path aggregation network (PAN) which improves the feature hierarchy by bottom-up path expansion shortening the information pathway for signals coming from the lower layers [92]. With this setup, the team achieved compelling results in the averaged mAP@.50:.95 metric for the three YOLOv8 model instances 8n, 8s, 8m at 0.943, 0.977 and 0.948, respectively. Sample detection results of the different bacterial strains by the custom

built and trained model are displayed in Figure 10. Additionally, successful detection of the life threatening *S. aureus* was at mAP50 > 98% [93]. The reliability and power of these predictions can play a major role in rapid diagnosis and correct medication dispersion, saving resources and protecting human health. Since YOLOv8 is a lighter real-time detector, it remains to be seen what difference would it play compared to the two-stage detectors under the same conditions.

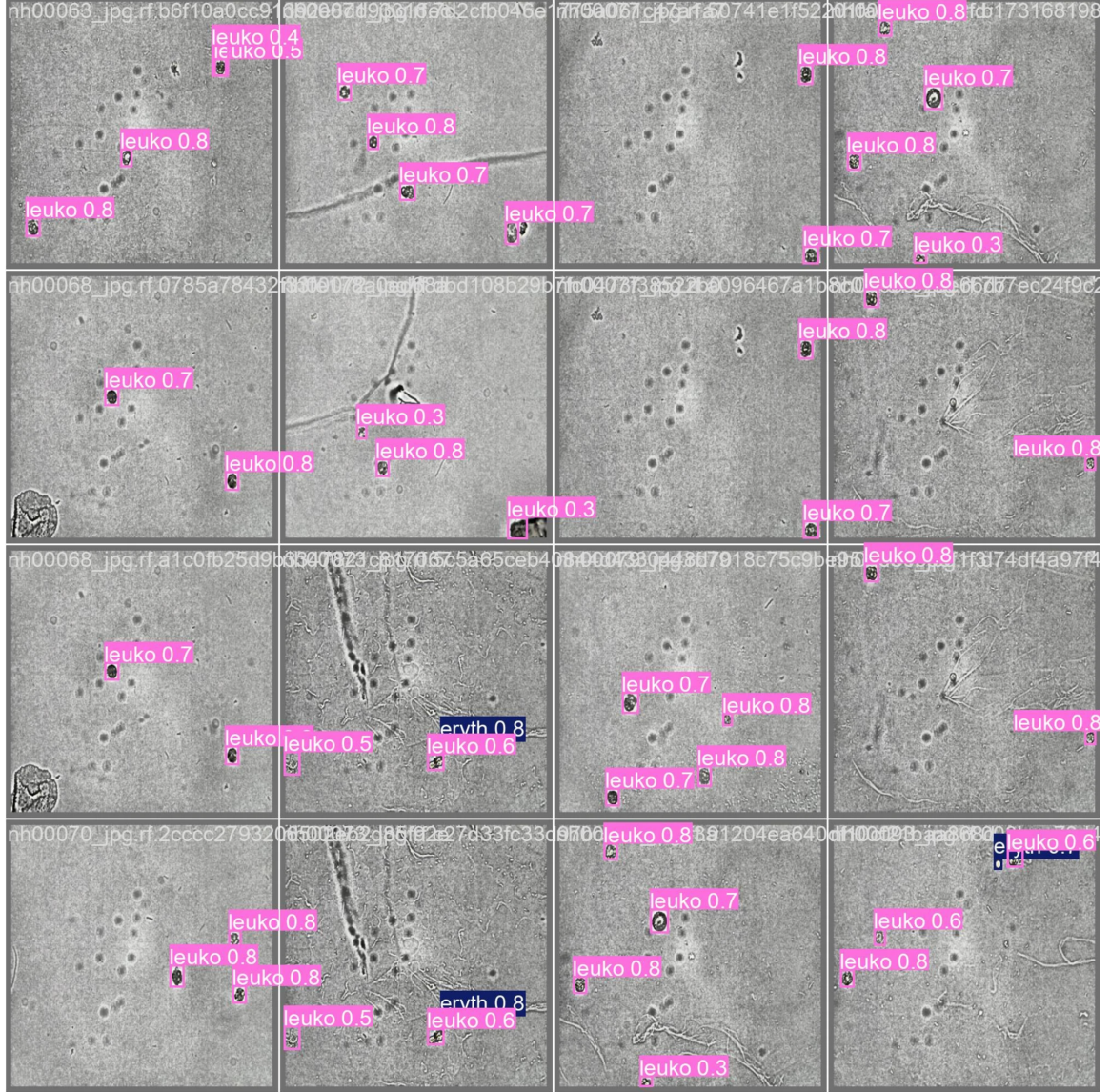


Figure 11: Predicted leukocytes in microscopic samples by the RF-DETR model. (Dahiya et al., 2025)

The advancements in transformer architectures allowed the following research team to create an image-based segmentation, detection and classification DNN-based framework for working with molecular, histological, microscopic and bright-field data. The study proposes CelloType as a path toward multitasking strategy, combining detection, segmentation and classification into a single stage, all while further increasing the performance. Similarly to

previous review, the team also leverages the use of Swin transformer as the backbone for multiscale feature extraction, which is then supplied to the DETR with improved de-noising anchor boxes (DINO) object detection module. As discussed, Swin backbone computes with shifted windows which limits self-attention to the overlap of local windows, but grants access to cross-window connection [82]. On the other hand, DINO is an end-to-end transformer detector utilizing a denoising method for improving the convergence of models sharing the DETR architecture. Supplying the transformer decoder with extra ground-truth labels and boxes, which have been affected by additional noise, is the main underlying principle. Additionally, DINO borrows the mechanism of query selection from Deformable-DETR, leading to initializing queries at improved positions [94]. The resulting architecture of CelloType is capable of handling variety of segmentation analysis tasks for both cellular and noncellular structures. The research team compared the results of the model with two other, namely Mesmer [95] and Cellpose2 [96]. The new model was able to significantly outperform the previous two with mAP of 0.56 in cell and 0.66 in nuclear segmentation tasks [97]. This novel multitasking model approach offers new insights into segmentation and classification in microbiology, however, the team also stated the cell type classification difficulty lingers and suggests further exploration on this topic.

Finally, another team based the development on an improved DINOv2 architecture along with deformable attention module during the construction of Roboflow (RF)-DETR and optimized it by the red fox optimization (RFO) algorithm. Such build allows for training on multiple resolutions and provides means for transfer learning into another field. However, it also requires greater time for hyperparameter tuning and the generalization capability of the model is limited in its default settings. The RFO is a mathematical model simulating the search for food, hunting and population emergence during the evasion from danger. The global and local optimization processes with added reproduction system is at the core of this algorithm [98]. The microscopic dataset the team utilized for train, test and validation splits was created at Roboflow site for annotation and model training [99]. Detection of small and medium objects, the leukocytes and erythrocytes, was the key focus of this experiment. Consequently, the team managed to achieve improvements in the mAP@0.5 detection metric, compared to the baseline models. Resulting 47% for small and 78% for medium sized objects, which is an improvement of 47% and 24% over the baseline model per each size. The predicted leukocytes in microscopic urine samples are shown in Figure 11, as predicted by the custom model. Nonetheless, the team reported the detection precision for larger objects remained dominated by the baseline models, suggesting that RFO specializes on small object detection only [100]. By demonstrating powerful improvements in small-object mAPs, the team has pushed the microscopic image analysis further. Despite the decreased performance on large-object mAPs, the model constructed in the study can be utilized for integration into a multi-modal system in the future.

2.6 Summary and Significance

The ML principles are part of an ever-growing arsenal of advanced algorithms. The chapter 2.4 explored the most recent research and deployments in the domain of general CV in terms of object detection and classification. It is worth mentioning, that applications in CV are not limited to detection but can also be exercised for high-level reasoning connected with spatio-temporal tasks, e.g., pattern recognition, interpretation, robotics, autonomous driving, facial recognition, scene understanding, visual questioning, optical flow, motion estimation or human-aligned perception among the few [101, 102, 103]. A summarized

list of models discussed in that section is provided in Table 1 for a concise review of the mAP@0.5 metric per each model, including the source article and use-case for reference.

Table 1: Reviewed models with reported mAP@0.5, source, and use case.

Model	mAP@0.5	Source	Use-case
AODSA-MPAVCP	78.83%	Adam et.al., 2025	Object detection (accessibility)
RT-DETR	88.0%	He et. al., 2025	Object detection (general)
YOLOv5	84.0%*	He et. al., 2025	Object detection (general)
YOLOv8	86.0%*	He et. al., 2025	Object detection (general)
DETR	83.0%*	He et. al., 2025	Object detection (general)
SSD	78.0%*	He et. al., 2025	Object detection (general)
HR-YOLO	79.8%	Zhang et. al., 2025	Foggy ADAS (road / traffic scenes)
Faster R-CNN	66.9%	Zhang et. al., 2025	Foggy ADAS (road / traffic scenes)
DETR	67.5%	Zhang et. al., 2025	Foggy ADAS (road / traffic scenes)
YOLOv8	81.0%	Wang et. al., 2025	Neuroblastoma histology (medicine)
YOLOv8-iRMB	82.0%	Wang et. al., 2025	Neuroblastoma histology (medicine)
YOLOv8-EVC	85.0%	Wang et. al., 2025	Neuroblastoma histology (medicine)
YOLOv8-IE	89.0%	Wang et. al., 2025	Neuroblastoma histology (medicine)
Faster R-CNN	82.0%*	Wang et. al., 2025	Neuroblastoma histology (medicine)

Notes: Values with * are inferred to normalize the view from comparative statements in the text (e.g., “X outperformed Y by 4%”) rather than explicitly reported as absolute mAP@0.5 reported in the chapter 2.4.

In alignment with this thesis, focusing on the aspects of localization, detection and correct assignment of pre-defined object classes in the realm of microscopic data is the central problem. Thus, the three most prominent architecture designs were selected and latest operational samples were described in the chapter 2.5. This chapter confirms the plausibility of utilizing these algorithms and provides real-world examples on construction and use for each. However, the summary of collected instances for each model reveals an amount of variability in precision results. This suggests more task-specific performance rather than a global generalizable metric. In fact, each of those studies is missing a direct comparison with a different family of models through the mAP metric, averaged anywhere between 0.5 and 0.9 IoU threshold. Furthermore, most studies assign the greatest value to the final models, without enough deep exploration into the cornerstone of machine learning, the source data. Consequently, granting the opportunity to investigate the impact of diverse algorithms for the same task as well as broadening the knowledge of the impact different pre-processing and augmentation steps might have on the end results. Additionally, the data challenges regarding the lack of available and validated microscopic datasets, defined in section 2.3, remain. Thus, providing the opening for contribution to the growing supply of re-usable open data sources. As part of this thesis, the imagery-data are annotated and compiled into various training datasets, depending on auxiliary data augmentations and versioning. Moreover, after the data are labeled they are also validated by an expert in the field and made available for future processing in accordance with the FAIR Data methodology. Hence, contributing to the expanding options for training on microscopic data. Finally, the development of additional DNN solutions provides new opportunities for transfer learning into other areas in microscopic image analysis and opens a way for future research and collaboration.

Table 2: Reviewed microscopy models with highest achieved mAP, source and use-case.

Model name	mAP	Source	Use-case
Faster R-CNN	0.493	Pawłowski et al., 2022	Microbial species
Cascade R-CNN	0.520	Pawłowski et al., 2022	Microbial species
YOLOv8n	0.943	Romphosri et al., 2024	Bacterial identification
YOLOv8s	0.977	Romphosri et al., 2024	Bacterial identification
YOLOv8m	0.948	Romphosri et al., 2024	Bacterial identification
DETR w/ DINO	0.660	Pang et al., 2025	Cellular/nuclear detection
DETR w/ DINOv2 and RFO	0.780	Dahiya et. al., 2025	Leukocyte detection

Exploration done in previous chapters into the ML and DNNs revealed the contemporary precision results in the fields of general CV as well as in the confines of microscopic imagery-data. Despite the large variance across the scores in Table 2, pertaining to the domain of microscopy, it serves as a guiding benchmark for this thesis. The aim remains to approximate or improve detection and classification accuracies of any one of the models, while focusing on the relevant real-world application in microscopic data analysis.

2.7 Contributions

While machine learning has driven significant advancements in macroscopic image analysis, the variability inherent to microscopic slide preparation frequently limits researchers to brittle, „one-time“ algorithmic solutions. Existing standard models continuously struggle with the detection and precise bounding of small to medium-sized biological objects scattered across complex backgrounds. To overcome these barriers, this thesis proposes a versatile, highly expandable domain-specific framework rather than a singular, isolated script.

The primary significance of this research lies in its holistic approach to the automation of pharmaceutical botany, offering the following core contributions to the field:

- **Comprehensive Architectural Benchmarking:** This study conducts a rigorous comparative analysis of the three most powerful contemporary object detection paradigms—single-stage (YOLOv10x), two-stage (Faster R-CNN), and transformer-based (DETR) architectures—establishing a definitive performance baseline for bright-field microscopic environments.
- **Fine-Grained Subclass Classification:** Moving beyond broad, generalized target detection, this framework successfully tackles the compounded challenge of hierarchical classification. It proves the models’ capability to reliably distinguish not only general botanical classes but highly specific, morphologically similar subclasses (e.g., distinguishing between diverse trichome structures and specialized crystalline druses).
- **Integration of Diagnostic Transparency (XAI):** To prevent the deployment of opaque „black-box“ systems, this work implements an Explainable AI pipeline. By integrating customized Grad-CAM extractions and spatial detection-density heatmaps, the study guarantees that algorithmic predictions are anchored to verifiable morphological features, fulfilling the transparency requirements of medical and pharmaceutical laboratories.

- **Publication of a FAIR-Compliant Bio-Image Dataset:** High-quality, expertly annotated microscopic datasets remain one of the most severe bottlenecks in the field. A highly significant contribution of this work is the curation, digital annotation, and optimization (v17) of the pharmaceutical botanical dataset. In adherence to the global FAIR (Findable, Accessible, Interoperable, and Reusable) data initiative, this dataset will be made publicly available to democratize future research and accelerate the development of specialized backbone architectures.
- **Foundation for Edge Deployment:** By utilizing state-of-the-art architectures and optimizing tensor resolutions down to 640×640 , this research successfully establishes the viability of real-time online inference tools. This guarantees that the proposed methodology is not confined to high-performance computing clusters, but can be seamlessly transitioned into practical, on-site laboratory deployment.

2.8 Research Goals and Questions

The overarching aim of this thesis is to bridge the gap between traditional pharmacognostic analysis and modern artificial intelligence by evaluating the efficacy of state-of-the-art deep learning architectures in the microscopic domain. While macroscopic plant identification has seen significant automation, the automated detection of highly imbalanced, complex microscopic botanical structures—such as those found in *Verbascum* and *Althaea officinalis*—remains an under-explored frontier.

Primary Research Goals

1. **Architectural Evaluation:** To systematically compare the performance, convergence stability, and statistical reliability of three distinct object detection paradigms: single-stage convolutional networks (YOLOv10x), two-stage region-proposal networks (Faster R-CNN), and vision transformers (DETR), when applied to high-resolution microscopic imagery.
2. **Dataset Optimization:** To establish a rigorous, empirically validated preprocessing and augmentation pipeline capable of mitigating the severe class imbalances and spatial biases inherent to manual microscopic slide preparation.
3. **Diagnostic Transparency:** To transition deep learning models from „black-box“ statistical calculators into verifiable diagnostic tools by deploying post-hoc Explainable AI (XAI) techniques, ensuring that the networks identify true biological morphologies rather than spurious background artifacts.

Research Questions

To fulfill these objectives, this research is guided by the following specific questions:

- **RQ1: Architectural Superiority:** How do single-stage, two-stage, and transformer-based object detection architectures compare in localizing and classifying microscopic pharmaceutical structures, and can one paradigm demonstrate statistically significant superiority (bounded by Wilson Confidence Intervals) in this highly specialized domain?
- **RQ2: Impact of Data Curation:** To what extent does strategic class-consolidation and targeted geometric augmentation improve model precision and recall compared

to raw, highly imbalanced datasets or datasets subjected to aggressive, generalized photometric distortions?

- **RQ3: Explainability and Feature Validation:** Can spatial detection-density heatmaps and gradient-based saliency mapping (Grad-CAM) effectively validate that a high-performing object detector is anchoring its confidence scores to true botanical features (e.g., cellular walls, trichomes, and druses) rather than operator-induced centering biases or slide illumination anomalies?

By addressing these questions, this study intends to provide a highly reproducible, empirically grounded framework for the future integration of computer vision systems in pharmaceutical botany laboratories.

2.9 Ethics and Reproducibility

As the application of artificial intelligence increasingly intersects with medical and pharmaceutical disciplines, addressing the ethical dimensions and ensuring the strict reproducibility of computational experiments are foundational academic imperatives.

The primary dataset utilized in this research consists exclusively of digitized microscopic bright-field images of botanical specimens, specifically *Althaea officinalis* and the *Verbascum* genus, cultivated and harvested at the Medicinal Plant Garden. Consequently, this study does not involve human subjects, animal testing, or Personally Identifiable Information (PII), thereby exempting it from institutional biomedical ethics board reviews regarding patient privacy.

From an algorithmic ethics perspective, the deployment of „black-box“ neural networks in quality control pipelines poses a risk of automation bias, where human operators defer to algorithmic decisions without understanding their basis. To ethically mitigate this risk, diagnostic transparency was established as a core pillar of this research. The integration of Explainable AI (XAI) techniques, specifically Grad-CAM and spatial heatmaps, ensures that the model’s feature-extraction rationale remains fully transparent and subject to human verification, aligning with the ethical guidelines for trustworthy AI in laboratory environments.

A pervasive challenge in contemporary deep learning research is the reproducibility crisis, often driven by undocumented hyperparameter tuning, proprietary datasets, and stochastic training variations. To guarantee the reproducibility of the empirical findings presented in this study, a rigorous standardization protocol was enforced:

- **Dataset Versioning:** All iterations of the dataset—including the baseline (v6), the heavily augmented (v13), and the finalized class-consolidated version (v17)—were strictly version-controlled within the Roboflow platform. The exact 87-8-4% dataset splits remain computationally locked, ensuring that future models can be evaluated against the exact same testing tensors.
- **Deterministic Hardware and Software Environments:** All computational experiments were executed within the standardized Google Colaboratory environment utilizing NVIDIA A100 and Tesla T4 hardware accelerators. The software stack was entirely built upon open-source frameworks with specific version dependencies, namely PyTorch for tensor operations, Ultralytics for the YOLOv10x architecture, Detectron2 for Faster R-CNN, and Hugging Face Transformers for DETR.

- **Hyperparameter Transparency:** All critical initialization parameters—including base learning rates, weight decay configurations (e.g., 3×10^{-4} for AdamW), cosine annealing schedules, bounding box IoU thresholds, and early stopping patience—have been explicitly documented in the respective architectural configuration sections.
- **Pre-trained Weight Initialization:** To avoid the chaotic divergence inherent to random weight initialization on small datasets, all models were initialized from publicly available, standardized Microsoft COCO pre-trained checkpoints, ensuring a mathematically identical starting point for the gradient descent optimization.

By transparently documenting the data provenance, architectural hyper-parameters, and exact computational infrastructure, this research provides a fully replicable methodological pipeline for future studies in automated pharmacognostic analysis.

3 State-of-the-Art

Establishing the scope of the current thesis is the main goal of this section and its subsections. This is achieved through the outlining of theoretical background for ML to specify the boundary from broad AI to narrow DNN architecture designs applied specifically in the field of CV. Consequently, the following subsections define concise terminology relating to the key concepts and features. First, the subchapter 3.1 provides a high level overview on the topic of ML, basic terminology, notation, schematics, core modules and training. Second, the subsequent chapters 3.2, 3.3 and 3.4 delve deeper into inner mechanics of the distinct architectural families along with strengths and weaknesses of the three selected network designs - i.e., Faster R-CNN, YOLO and DETR. Next, the subsection 3.5 details approaches to model assessment through the systematic use of metrics, benchmarks and touches on the issues of XAI. Finally, the subchapter 3.7 summarizes the critical concepts discussed throughout this section. Resulting in acquisition of elemental knowledge necessary for the subsequent experiments in the application section 4 of this paper.

3.1 Foundational Concepts

The first part presents an outline on the basic premise of ML, how it operates and what is the added value of its deployment. The introduction into the theoretical setting of supervised learning is next. Followed by the fundamental ML terminology - i.e., layers, activation and loss functions, optimization, convolution, residuals, attention mechanics, backbones, pipelines, augmentation and transfer learning. Hence, the following divisions aim to establish founding theory on the general ML components. This knowledge is crucial for later chapters which detail the individual architecture designs and metrics.

3.1.1 What is Machine Learning

The need for large-scale datacenters to generate, store and then process data has already passed. Today, the immense amount of data is generated every second from rudimentary activities like listening to music, walking, driving or utilizing any online services, e.g., social media, shopping, video sharing, streaming, gaming, news channels, conferences, logistics, or writing a simple blog post. The data produced via such activities are immediately processed and turned into predictions to customize any future content of each individual. This is no longer a by-product but a standardized feature of most online services [104, 105]. How can a store owner evaluate whether a product is going to bring them a profit? The patterns based on the past sales, seasons, trends, or crises might. Thus, collecting data and performing analysis can bring new observations which create the base for future decisions for the current business, as seen in Figure 12.

Leveraging the power and versatility of modern computers, many problems can be transformed into an algorithm. Any sequence of individual steps or mathematical rules producing outputs from inputs defines the simplest algorithm. With similar intuition, when an algorithm is capable of recognizing patterns in a dynamic environment, converting them into predictions while also learning from those results, it can effectively simulate intelligence. Hence, the term AI forms the broadest spectrum of such algorithms. Although, one problem can be solved with more than one algorithm, in the world of AI, more problems can also be solved by the same algorithm. Thus, leading to a challenge of conserving the resources - i.e., memory size, processing time, energy consumption and number of instructions required. Nevertheless, solving certain tasks has become intuitive for humans, thus, problematic to



Figure 12: Data analytics being used for a data-driven decision making. (Kelleher, 2020)

describe with a fixed set of rules. Approximating and assigning the correct importance (the weight) per each input to deliver desired outputs forms the core of AI algorithms. Essentially, this approach allows the program (the machine) to estimate the impact of these initial weights on the final product on its own, based on the large volumes of data (the experience) it has access to (the learning). This learning is a way the computer can simulate the intuition to solve the tasks a human has difficulty describing [104, 106]. Learning, much like in humans, is about exploring a system and identifying its constituents. It is not a perfect process but rather a continuous approximation, as is the ML.

In order to develop a successful mining operation, the soil needs to be mined to localize sought ore veins. Once these deposits are identified, a mine is constructed around these areas to maximize the profit. As these veins are mined, the optimal location can change over time, requiring adjustments to these locations. Similarly in ML, mining large datasets, denoted as *data-mining*, devise the critical path to find the optimal set of weights. Once found, these weights define the initial cornerstones of an effective model. However, for the most stable stream of accurate predictions, weights might have to be adjusted, moved, or some even removed over time. Additionally, the hyperparameter optimization (HPO) influences the learning process from the outside. If weights are internally learned by the machine then hyperparameters are externally set by a human. Furthermore, in recent DNN architectures with dozens to hundreds hyperparameters available, the goal is for the machine to automatically find the optimal values to optimize performance of the network. Among the most significant contributions of the HPO in automated ML is minimizing human interference, enhancing the adaptation capabilities of the model and increasing the reproducibility of the results. [106, 107] Resuming the former excavation analogy, the HPO resembles the situational adjustment of the tools employed. For initial prospecting, durable and portable drills to extract samples are required, while precision ore extraction requires specialized diamond-tipped drills for cutting around the ore-veins without damaging the material.

3.1.2 Why Machine Learning

With the constant stream of data and modern databases being able to hold terabytes of tables with millions of rows [108], the task of analysis such data can challenge hundreds of human analysts. This is especially true for intuitive trend investigation, where data can lead to one end, but the multiple relationships affecting the trend remain obscured or seem unrelated. This perception of data invites efforts to create an explicit algorithm, explaining these indicators through a rational list of steps. However, this is where the limits

of what can be converted into an algorithm are met. Additionally, with such large volumes of data available, the next logical course of action is to utilize the data for resourceful decision making. However, how can data become resourceful in the absence of any means for interpretation?

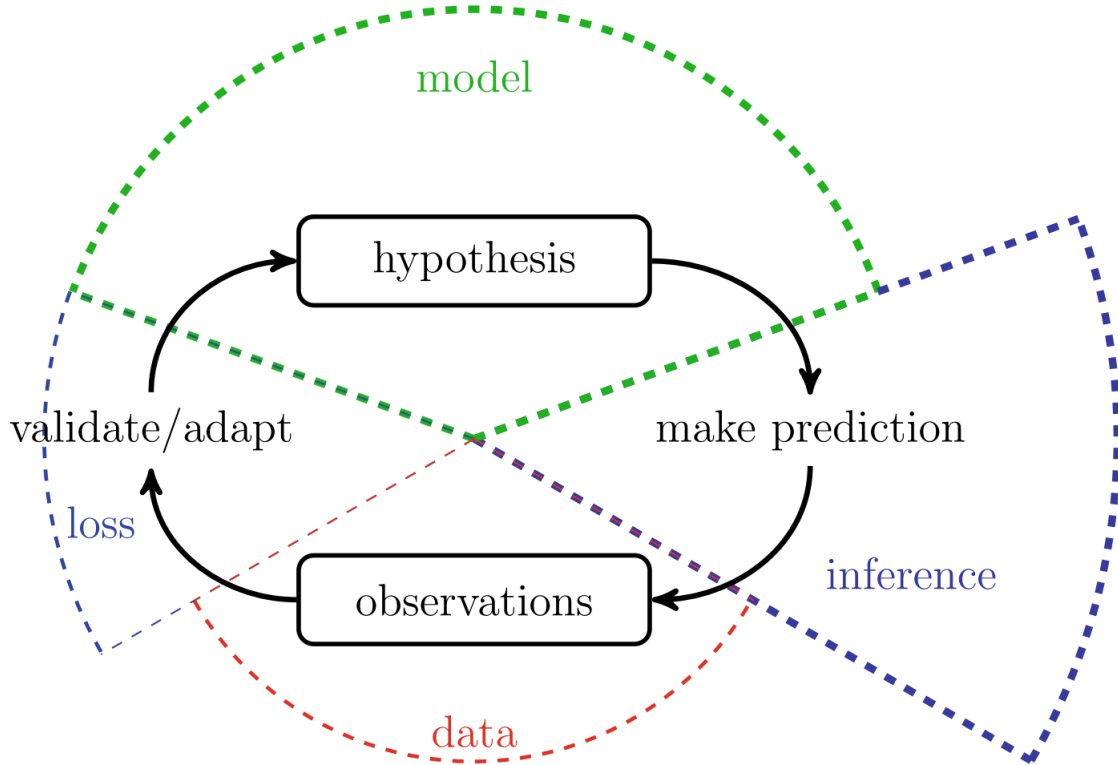


Figure 13: Main components of ML: data, model and loss. Based on the continuous cycles of trial-and-error, each inference produces loss which is then used in subsequent runs, producing improvements of the model. (Jung, 2022)

The answer is experience. ML allows to generalize from sample data (the experience) to create an estimation. Mathematical methods of iterating an increasingly more accurate approximation through countless of cycles bypasses the need for an explicit algorithm. Furthermore, combination of data, model and loss executes a fundamental „trial-and-error“ convention to refine the hypotheses about the data with every iteration. The specific issue of creating data driven decisions out of multilayered data is made possible with the use of DL which is capable of mapping complex inputs to decisions. Such extraction of patterns and functions is performed automatically and is especially significant in fields of vision, speech or language tasks where hand-coded programs fail. By learning the hierarchical feature representation from the vast amount of complex data, the DL generalizes from examples and relies on indicator features to create breakthrough solutions [104, 106, 109, 105]. Connecting to the ore deposit analogy, the geological research [110] has confirmed that various chemical associations between haloes of elements such as arsenic (As), antimony (Sb), bismuth (Bi) and sometimes copper (Cu) or tellurium (Te) are often correlated with discovery of golden ore veins. Thus, such particles are dubbed *pathfinder* elements as the steady rise of concentration correspond to an increased chance of nearby golden deposit

discovery. This is similar to a DL algorithm data-mining through extensive databases to determine which indicators have the greatest impact on the presence of sought prediction.

To conclude, ML provides the most substantial benefit in complex tasks with the lack of hand-coded algorithms but abundance of collected data. The goal of ML is to learn, generalize and predict or identify. This is done through learning on the available data, with subsequent inferences performed on data never seen before. Repeated validation and refinement is based on the loss incurred during these inferences. This can be seen in the concept diagram in Figure 13, which describes how comparing inferences to data generates loss, which is then employed for revising the hypothesis (the model) in each run. Furthermore, the latest line of models is based on the DL discipline which aims at multi-level feature extraction for simple, local and higher layers of patterns. When combined, these layers are capable of learning and delivering progressively more abstract results. Thus, being most often assigned to areas of vision, speech and language where connections between explicit and underlying meaning can be ambiguous to standard procedural programming. Finally, the scalability of practical and commercial solutions can be achieved by automated ML by delivering optimized parameters for models deployable by non-experts or students.

3.1.3 Supervised Learning Framework

Definition of the correct output per given input is done solely by a human supervisor. This constitutes the core difference of supervised ML from any other type. The goal is to teach the model the ground-truth through assignment of classes or providing the sample data. By experiencing data through the lenses of desired features, labels and targets, the algorithm can learn to recognize these targets in previously unseen datasets based on the measurements associated with the training samples, seen in diagram Figure 14. Additionally, while learning from data points, the model has access to all the available features and can assess which of these features contributes to the target the most. Thus, the most logical route is to discard any non-relevant features and focus the attention only on those which pass a certain threshold of relevance. By optimizing the learning speed, the loss against the labels is being minimized as well [104, 106, 109, 105]. As mentioned, supervised approach simulates the experience gained from recurring exposure to the training data, e.g., children familiarizing with the names of farm animals. After the initial reveal of unknown species, most children will trigger false positives trying to classify the new beasts. However, repeated display of animals (data points) leads to the children identifying the unique features for each one. Thus, significantly improving the perception of the child (the model), leading to considerable identification precision (the inference) on previously unseen animals.

For instance, if the aim of a model is to predict whether a person will successfully graduate, the training dataset might carry millions of rows for past people who graduated and also those who did not. Each data point $\mathbf{x} = [x_1 \ x_2]^T$, represents an outcome where the desired features are defined as the number of hours of studying per day (x_1) and the averaged study results thus far (x_2). Relative to the sample diagram in Figure 14, the $\mathbf{x}$ represents the target feature and (x_1) and (x_2) are the descriptive features. Next, the optimized model weights $\theta = [\theta_0 \ \theta_1 \ \theta_2]^T$ must be determined. As discussed in earlier sections, a vast amount of data is used to find the optimal values for the $[\theta_0 \ \theta_1 \ \theta_2]$ parameters through learning and this is connected to the selection of the optimal ML algorithm for the given task, seen in Figure 14, where the algorithm receives the training data and outputs the predictive model. The complete model $h(\mathbf{x}, \theta)$ where each internal weight θ_n corresponds

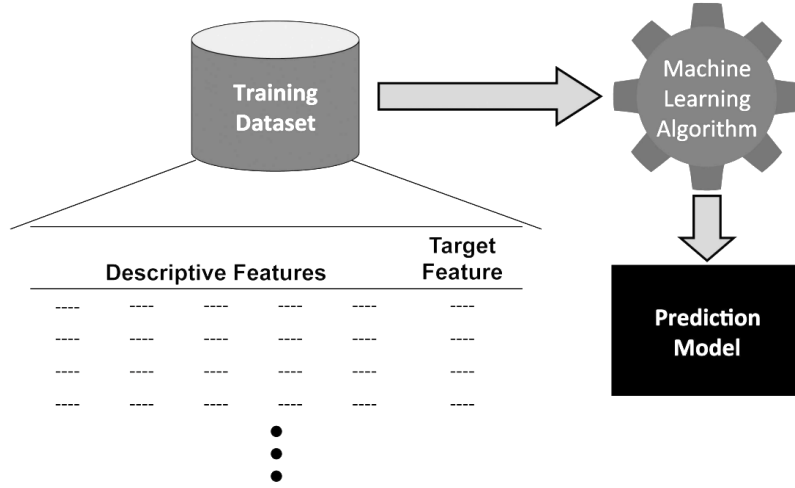


Figure 14: Data being used to train a new predictive model. (Kelleher, 2020)

to each selected feature x_{n+1} denoting the importance of the specific value. Exception is θ_0 which is referred to as *bias*, and signifies the y-intercept - i.e., the baseline value if all other input features equal zero. The bias is used for flexibility through shifting of the entire model in the feature space, accounting for unrelated patterns. Let $\mathbf{x} = (x_1, x_2)^\top \in \mathbb{R}^2$ denote the input feature vector from the dataset with the model parameters collected into the augmented vectors:

$$\tilde{\mathbf{x}} := (1, x_1, x_2)^\top \in \mathbb{R}^3, \quad \tilde{\theta} := (\theta_0, \theta_1, \theta_2)^\top \in \mathbb{R}^3. \quad (1)$$

Then, the linear classifier $h : \mathbb{R}^2 \rightarrow \{-1, 1\}$ is defined as

$$h(\mathbf{x}; \tilde{\theta}) = \text{sign}(\tilde{\theta}^\top \tilde{\mathbf{x}}), \quad \text{with} \quad \text{sign}(z) = \begin{cases} -1, & z < 0, \\ 1, & z \geq 0, \end{cases} \quad (2)$$

where the bias term θ_0 is included in $\tilde{\theta}$. Equation 2 defines a linear decision boundary $\{\mathbf{x} : \tilde{\theta}^\top \tilde{\mathbf{x}} = 0\}$, where the dot product of $\tilde{\theta}^\top$ and $\tilde{\mathbf{x}}$ is evaluated with the sign function, delivering the prediction whether the person is likely to graduate 1, or not -1, based on the time spent studying and their study average. Such approach is typical for a logistic regression, where upon crossing a pre-defined threshold, the data points are automatically separated into two distinct levels. From the geometrical perspective, this is an elementary linear classifier, dividing the 2D plane into half and sorting data points into two categories. Even though such model is convenient for features defined by linear relationship, it is severely lacking the ability to return valid results for any non-linear dependence. This limits the ability of the model to learn complex feature connections, thus, failing on predictions and classifications [104, 105, 111]. The classical and the original example of non-linear relationship is the XOR operation. Because of how XOR is implemented, it is impossible to separate $[0;1]$ and $[1;0]$ from $[1;1]$ and $[0;0]$ coordinates with a linear classifier.

An example of another non-linear relationship is observed every day on earth, the daylight. As the sunrise and sunset are determined by the geometry and the Earth-Sun relation, the elevation of the Sun over the year depends on sines, cosines and inverse trigonometric functions of latitude and the day of year. If the only available data are measurements of

celestial bodies in question, with no visual confirmation, then the prediction of the daylight duration for any given day of the year, is non-linear.

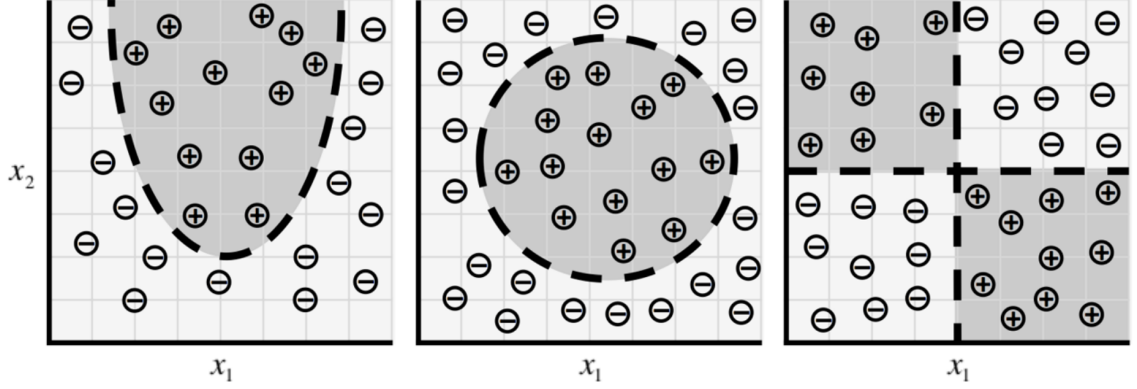


Figure 15: The geometrical representation of perceptrons with non-linear boundaries. From left: parabolic, circular and piecewise or step-like boundary. (Buduma, 2022)

Therefore, past research has focused on development of non-linear classifiers which would be capable of separating data points into two classes through different boundaries. The figure 15 exhibits the samples of non-linear classifiers used to separate data which a simple logistic regression would struggle to achieve. The major problems arise in attempts to modulate human-like abilities and senses, e.g., vision, hearing and speech. Such complex functions are exceedingly more dimensional and non-linear to capture the true underlying relationship between the relevant features. Consequently, the research has moved passed the support vector machines (SVMs) with non-linear kernels and decision trees (DTs) to neural networks (NNs) [104, 111]. Even though the methods of SVMs and DTs are still frequently applied, the main focus has shifted towards NNs. The main reasons are the increase in computational power and storage capacity, both combined, allowing for construction of robust multi-layer solutions with generalization capabilities. Worth noting is that NNs are not limited to application under supervised learning, but are utilized also in unsupervised, semi-supervised, weakly-supervised and transfer learning methods of ML. However, to maintain the scope of this thesis the other learning approaches are not further developed in this work.

3.1.4 Deep Neural Networks

As a specialized subfield of AI the *deep learning* is based on the NNs with many subsequent (the hidden) layers between the input data and output prediction, hence, the *deep* NNs. The neuron, being the basic unit formulating cognition in humans, became the source inspiration for development of algorithms simulating cognition in computers. With 86.1 +/- 8.1 billion neurons in human brain, it is clear these units are not connected sequentially. Rather, interconnections between neural layers upon layers of brain cells creating the neural networks of intricacy difficult to imagine [112, 113]. Similarly, the artificial neural networks are composed of certain number of layers with varying layouts depending on options in the architecture design of the network. Generally, these are the input and output layers, with a set of hidden layers for inner data processing. When the data is handled in a forwarding manner between the layers, it is denoted as the feed-forward neural network

(FNN) [105, 111, 114]. Another type is backpropagation, where some neurons of higher layers are sent back to the lower levels for consideration of the initial output. The simple representation of a FNN is in Figure 16.

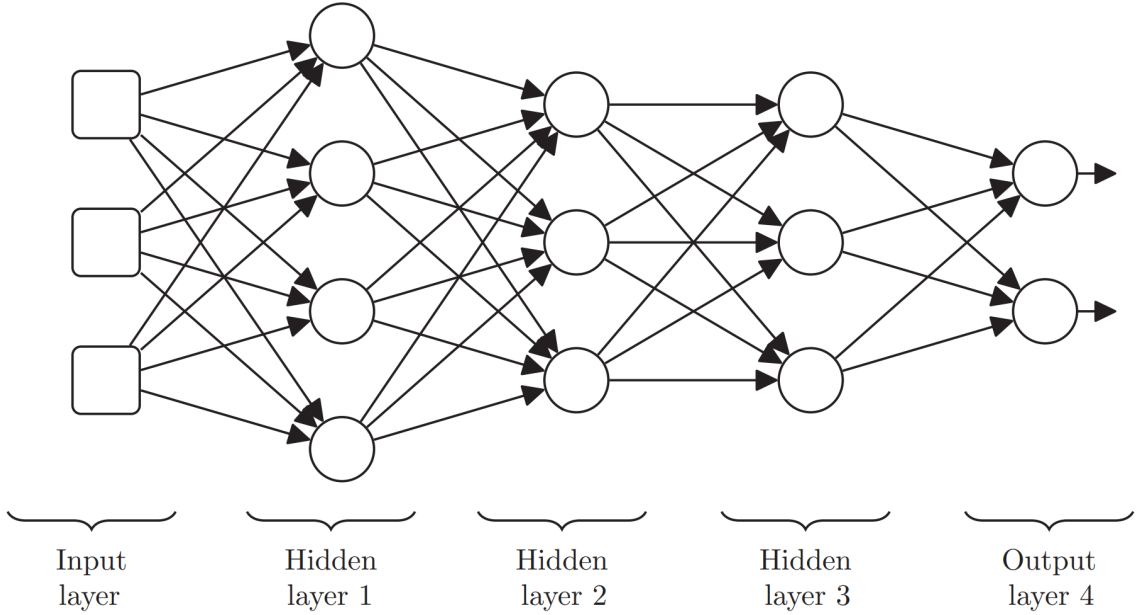


Figure 16: Example structure of an artificial FNN. (Kelleher, 2019)

The analysis occurs much like in the human brain, for instance, when the brain first receives sensory data from the eyes functioning as the visual cortex exciting the input layers. The processed information continues through the follow up sections, leading to the output layers in the form of conceptualizing the message. This leads to a realization that there is a cat and not a car, standing in front of the human [111, 115]. Multi-modality, in this example, would include touch of the fur or hearing the purr. Leading to more activations along different processing-centers (the models) of the brain and combining the results into the final definition of reality.

The critical path from the input layers to the output layers is represented by the hidden layers. Creating the core which determines the successful solution of the problem. Furthermore, the hidden layers serve as an automated feature detector and evaluator. Hence, it is possible to examine these structures to derive meaningful features detected by the DNN. Analyzing these layers becomes critical for ML models utilized for decision making, where reasoning becomes a necessary part of the verdict. Methods developed for such purposes are categorized under Explainable AI (XAI). The main product of these systems is employed in legal, psychological and ML validation requirements during research, development and application [109]. This is all made possible through the hidden layers, which often have smaller number of neurons between layers than the input layer to pressure the entire network to learn the condensed representation of the original information. However, the main difficulty lies in interpreting the extracted information to the human-understandable format [111] The explainability is discussed further in the following section 3.6.

3.1.5 Convolutional Networks

Processing image data requires a grid-like topology to represent the 2-D grid of pixels and means to process it. The Convolutional Neural Networks (CNN) are utilizing the specialized linear mathematical operation named convolution. In this operation the x is the input at time t , where both x and t are real numbers. This means x can return different values based on the time t and the more recent measurement are more relevant. Thus, the weighted function of $w(a)$ is applied, where a is the age of the measurement. This creates a function s which yields a smooth estimate [106, 116].

$$s(t) = \int x(a)w(t-a)da. \quad (3)$$

This is the fundamental definition of convolution operation, which is commonly abbreviated with an asterisk, downsizing the formula to:

$$s(t) = (x * w)(t). \quad (4)$$

In CNN design the argument x is the input while w is denoted as the *kernel* and the output as the *feature map*. However, time is most often measured in the discrete and concrete values, such as seconds. Sample diagram from the LeNet architecture is displayed in the Figure 17. This implies that measurements of $\mathbf{x}$ and w exist only on the integer values of t and thus allows to be define with the following sum:

$$s(t) = (x * w)(t) = \sum_{a=-\infty}^{\infty} x(a)w(t-a). \quad (5)$$

Apart from the ability to work with inputs of variable size the convolution brings three core pillars into the ML. These are:

- Sparse Interactions
- Parameter Sharing
- Equivariant Representations

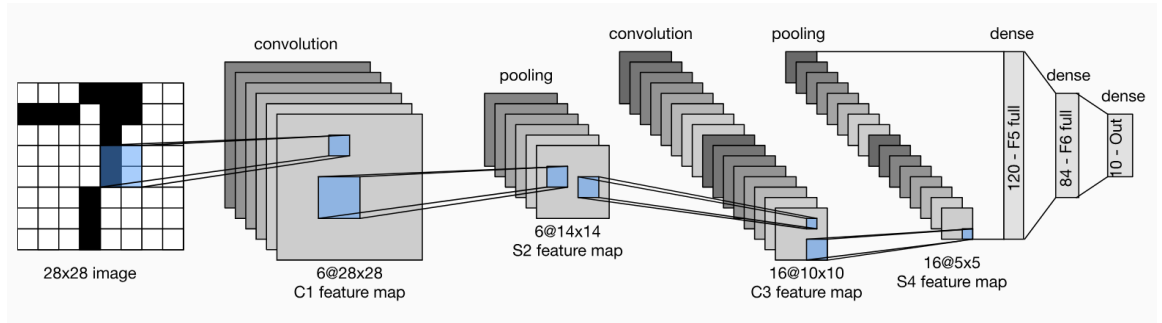


Figure 17: Sample of data flow in LeNet Convolutional Neural Network. (Zhang, 2023)

First, to define sparse interactions is to contrast network layer usage with the traditional neural networks. There, every input element interacts with every output element, using multiplication of parameters with every single parameter. However, the convolutional

networks apply smaller kernels than the input, thus take full advantage of *sparse interactions*. This allows to detect meaningful patterns in small areas such as the edges of objects and frames. Smaller kernels directly translate to fewer parameters, reducing the memory needs while increasing efficiency. Next, the parameter sharing indicates that the value of one weight parameter, connected with one input, influences a value of another weight in a different place of the network. Finally, the equivariance can be described that a function $f(x)$ is equivariant to a function g if $f(g(x)) = g(f(x))$. This means that whether we perform transformation through the function g and then execute the convolution operation, the result is going to be the same if we swap the order of these activities. When processing images, this aspect creates a 2D map describing the features inside the input. It can also be observed for time-series data where convolution creates a type of timeline for each feature presented in the input. Hence, moving an object in the input the mapping of the main features will be translated in the output as well. This is crucial for detecting the same edges in different regions of one image, for instance [106, 116, 114].

3.2 Two-stage Faster R-CNN

One of the first computer vision detectors created through the neural networks is the R-CNN, which was first introduced by Girshick, Donahue et al. (2014). Through the selective search algorithm, this model first extracts a large number of proposals. Each of these regions is transformed and warped into a smaller input sized image, labeling bounding boxes and classes. Next step is performing a forward propagation of the data through either through AlexNet or a VGG style net, similar to previously mentioned the VGG19. Finally, the architecture employs the support vector machines (SVM) as the final confirmation classifier. Even though the process take advantage of pre-trained CNNs, for a thousand of region proposals, it requires a thousand of CNNs for predictions. This slows the model down.[71, 106, 117] The R-CNN architecture diagram can be seen in Figure 18.

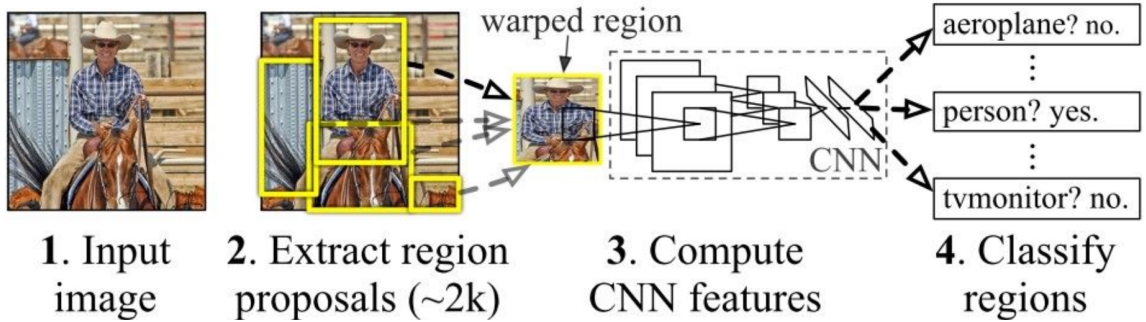


Figure 18: The original R-CNN architecture diagram. (Girshick, 2015)

Later, an innovation by Girshick (2015) was introduced in the Fast R-CNN model. Here, the CNN and region extraction sections are updated to perform the forward propagation on the entire image and not on thousands of proposals independently. Which previously led to using CNNs once on each region and then discarding them. Furthermore, the SVM classifier is replaced by fully connected layers which are utilized for computing the bounding boxes and classes. This increased the training and testing times through the re-use of the CNN calculations [71, 106]. Finally, the Faster R-CNN emerged through the work of Ren, He et al. (2015). Further increasing the speed of inference by replacing the slow selective

search section with a separate convolutional RPN. This section computes the convolutional features and roughly estimates the anchor box locations for variable number of objects. Final processing is performed by the Fast R-CNN head, where the object detections are sorted and merged via non-maximum suppression [116, 118, 117].

3.3 Single-stage YOLO

Unlike two-stage detectors that rely on sequential region proposals and classification, the “You Only Look Once” (YOLO) architecture, introduced by Redmon et al. [119], revolutionized object detection by framing it as a single regression problem. YOLO predicts bounding box coordinates and class probabilities directly from image pixels in one forward pass, drastically increasing inference speed [119]. The input image is divided into an $S \times S$ grid, where the cell containing the object’s center is solely responsible for predicting its bounding boxes and conditional class probabilities. Modern YOLO architectures have evolved into a stabilized tripartite structure [120]:

- **Backbone:** Extracts multi-scale feature maps. Contemporary models utilize Cross Stage Partial Networks (CSPNet) [121] to optimize gradient flow and reduce computational redundancy.
- **Neck:** Aggregates features using Path Aggregation Networks (PANet) [122] and Feature Pyramid Networks (FPN). This bottom-up and top-down fusion is essential for detecting objects of varying sizes, a frequent requirement in microscopic image analysis [123].
- **Head:** Executes final predictions. Recent iterations favor anchor-free detection, predicting object center points and dimensions directly to better accommodate irregular morphologies.

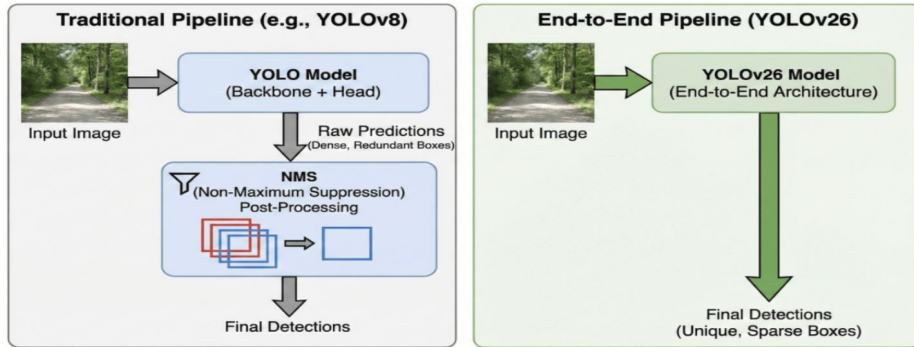


Figure 19: Comparison of NMS-based (LEFT) vs. NMS-free (RIGHT) architecture. (Chakrabarty, 2026)

Historically, YOLO models relied on Non-Maximum Suppression (NMS) to filter redundant bounding boxes generated by one-to-many label assignments. This post-processing step introduced variable latency, hindering true end-to-end deployment.

The YOLOv10 architecture addresses this bottleneck by introducing a consistent dual assignment strategy for NMS-free training [25, 124], seen in the Figure 19. It employs both

a one-to-many branch for dense supervision and a one-to-one matching branch. During inference, the one-to-many head is discarded, enabling definitive predictions without NMS overhead. Furthermore, YOLOv10 utilizes a lightweight classification head, spatial-channel decoupled downsampling, and partial self-attention modules to significantly reduce floating-point operations (FLOPs) while maintaining high mean average precision (mAP) [25, 124]. This balance of low latency and high accuracy makes it optimal for the rapid analysis of dense microscopic datasets.

3.4 Transformer-based DETR

While YOLO and its convolutional counterparts dominate real-time detection, they inherently rely on localized receptive fields and heuristic post-processing algorithms. To fundamentally rethink this paradigm, Carion et. al. [80] introduced the DEtection TRansformer (DETR), which reformulates object detection as a direct set prediction problem, presented in the Figure 20. By discarding hand-crafted components such as anchor boxes and Non-Maximum Suppression (NMS), DETR offers a streamlined, end-to-end architecture that leverages global contextual reasoning [80, 125].

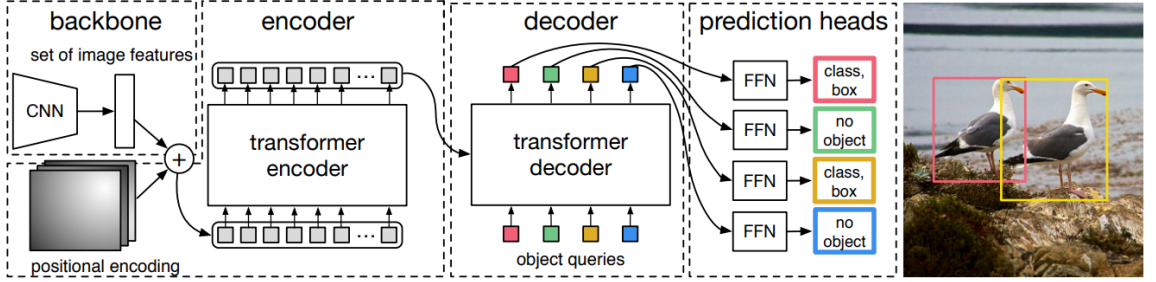


Figure 20: Architecture of the DEtection TRansformer (DETR). A standard CNN extracts feature maps, which are flattened and supplemented with positional encodings before passing into a transformer encoder-decoder sequence. The decoder utilizes a fixed number of learned object queries to directly output a set of bounding box predictions and class probabilities. (Carion et al, 2020)

The canonical DETR framework operates through a sequential pipeline diverging significantly from traditional CNNs [125, 125]:

- **Backbone and Encoder:** A conventional CNN backbone extracts a lower-resolution feature map, which is flattened and augmented with positional encodings. This sequence is processed by a Transformer encoder, utilizing self-attention mechanisms to learn global spatial dependencies and relationships across the entire image.
- **Decoder and Object Queries:** The Transformer decoder takes a fixed number of learned positional embeddings, termed “object queries.” Through multi-head cross-attention, these queries interact with the encoder’s output, iteratively refining their representations to focus on specific target regions.
- **Bipartite Matching Loss:** The network directly predicts a fixed-size set of bounding boxes and class probabilities. During training, the Hungarian algorithm forces a unique, one-to-one bipartite matching between the predicted and ground-truth objects, penalizing both incorrect classifications and bounding box deviations simultaneously [80].

Despite its conceptual elegance, the original DETR suffered from exceedingly slow training convergence and poor localization of small objects—a critical limitation when analyzing dense microscopic datasets. The computational complexity of the global self-attention mechanism scaled quadratically with the spatial size of the feature maps, restricting the use of high-resolution inputs necessary for detecting minute structures [125].

To overcome this, Deformable DETR [126] introduced deformable attention modules that attend only to a small set of key sampling points around a reference, significantly reducing computational overhead and enabling the processing of multi-scale feature maps. Building upon this, the Real-Time DEtection TRansformer (RT-DETR) by Lv et al. [127] successfully bridged the gap between transformer accuracy and YOLO-level latency. RT-DETR employs an efficient hybrid encoder that decouples intra-scale interactions from cross-scale fusion, alongside an uncertainty-minimal, IoU-aware query selection mechanism. This ensures high-quality initial queries and allows for dynamic adjustment of decoder layers to balance inference speed and precision [127].

The evolution of these architectures has made them exceptionally potent for complex microscopic image analysis. In domains where targets such as cellular structures, biological tissues, or botanical extracts exhibit high density and variable morphology, the global receptive field of transformer-based models excels at distinguishing targets from heavily textured backgrounds. Recent adaptations of the DETR paradigm in microscopy demonstrate superior performance in isolating small, irregular features without the computational bottleneck of NMS [100], solidifying the architecture as a highly robust alternative to purely convolutional networks.

3.5 Evaluation Protocols

To rigorously assess and compare the performance of the proposed deep neural network architectures—YOLOv10x, Faster R-CNN, and DETR—a standardized set of evaluation metrics and protocols is required. In the domain of object detection, particularly when analyzing complex microscopic imagery, relying on a single metric can obscure model biases, such as poor localization or an inability to detect targets of a specific size. Therefore, this study adopts a comprehensive evaluation framework based on the established Microsoft COCO metric standards [128], augmented with statistical confidence intervals to quantify prediction reliability.

3.5.1 Fundamental Metrics: IoU, Precision, and Recall

The foundational criterion for determining a correct detection is the Intersection over Union (IoU) [129, 130]. IoU measures the spatial overlap between the predicted bounding box (B_p) and the ground-truth bounding box (B_{gt}), calculated as the ratio of their intersection to their union:

$$IoU = \frac{Area(B_p \cap B_{gt})}{Area(B_p \cup B_{gt})} \quad (6)$$

A detection is classified as a True Positive (TP) if the calculated IoU exceeds a predefined threshold (commonly $\alpha = 0.50$). Conversely, detections falling below this threshold are marked as False Positives (FP), and ground-truth objects without a corresponding bounding box are recorded as False Negatives (FN).

These classifications directly feed into the calculation of *Precision* and *Recall*. Precision measures the exactness of the model (the proportion of positive identifications that were

actually correct), while Recall measures its completeness (the proportion of actual positives that were correctly identified) [130]:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN} \quad (7)$$

3.5.2 Average Precision (AP) and Mean Average Precision (mAP)

Because Precision and Recall fluctuate depending on the confidence threshold applied to the model’s predictions, the Precision-Recall (PR) curve is utilized to visualize this trade-off. The Average Precision (AP) encapsulates the area under this PR curve for a single class [130]. For multi-class evaluations, the Mean Average Precision (mAP) is computed by averaging the AP across all N classes [129]:

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (8)$$

This study primarily reports two standard variations of this metric [128]:

- **mAP@0.50:** The mean Average Precision evaluated at a single, permissive IoU threshold of 0.50. This metric is generally accepted as the primary indicator of successful object localization and classification.
- **mAP@0.50:0.95:** A stricter metric that averages the mAP evaluated across 10 IoU thresholds, ranging from 0.50 to 0.95 in steps of 0.05. This rigorously penalizes models with poor bounding box regression.

3.5.3 Size-Specific Metrics and Average Recall

Given the morphological diversity inherent to microscopic botanical samples, size-specific metrics are deployed to isolate performance bottlenecks. Detections are segregated into small ($area < 32^2$ pixels), medium ($32^2 \leq area \leq 96^2$ pixels), and large targets ($area > 96^2$ pixels) [128]. Furthermore, to assess the maximum detection capacity of the architectures before NMS or heuristic filtering culls the results, the Average Recall (AR) is measured given a fixed maximum number of detections per image (e.g., AR@1, AR@10, and AR@100).

3.5.4 Statistical Grounding: Wilson Confidence Intervals

Point estimates such as mAP or overall Precision lack insight into the variance and uncertainty of the model’s predictions, particularly when dealing with highly imbalanced datasets or classes with low instance counts. To address this, 95% Wilson Score Confidence Intervals (CIs) [131] are calculated for the reported metrics.

Unlike the standard Wald interval, the Wilson interval is asymmetric and does not assume a normal distribution, making it exceptionally robust for extreme probabilities (near 0 or 1) and small sample sizes [131]. By bounding the metrics within these intervals, the evaluation protocol accurately reflects the true statistical uncertainty of the detections, allowing for a highly objective comparison of generalization capabilities across the YOLOv10x, Faster R-CNN, and DETR frameworks. While the Wilson Score Interval provides robust statistical grounding, its application in object detection necessitates an understanding of its inherent advantages and constraints [132]:

- **Strengths:** The interval delivers superior coverage compared to standard Wald intervals, maintaining high accuracy even with small sample sizes (n) and extreme probability values (p) [133]. Furthermore, it enforces logical boundaries strictly within the $[0, 1]$ range and produces asymmetric limits that correctly reflect skewed statistical uncertainty [132]. Finally, it provides a computationally efficient, reproducible method that scales reliably across severe class imbalances without the computational overhead of bootstrap sampling.
- **Limitations:** Conversely, the metric relies on the mathematical assumption of independent and identically distributed (i.i.d.) detections, thereby ignoring hierarchical variance and image-level grouping [134]. Because it does not account for broader model or dataset variability, it may yield overly narrow, optimistic intervals in complex object detection environments. Additionally, while highly accurate for simple proportions, it serves only as an approximation for area-under-curve metrics such as mAP and lacks inherent corrections for multiple class comparisons.

3.6 Explainability

To address the inherent black-box nature of deep neural networks and validate the morphological features driving the model’s predictions, Explainable AI (XAI) techniques were integrated into the evaluation pipeline [135]. In the context of medical and microscopic image analysis, where trust and transparency are paramount, XAI provides critical verification that models are focusing on biologically relevant structures rather than spurious background artifacts [136]. This study specifically extracts spatial and saliency maps from the best-performing architecture, YOLOv10x, to provide human-interpretable verification of the detection outputs.

3.6.1 Detection-Density Heatmaps

To analyze the spatial distribution and potential localization biases of the model, two variations of detection-density heatmaps were employed: center-based and distributed mass over pixels.

- **Center-only Heatmap:** This method increments single pixel values corresponding to the exact center coordinates of each predicted bounding box. It is computationally lightweight and provides a rapid assessment of global detection clusters and potential dataset localization bias.
- **Distributed Mass Heatmap:** This approach spreads the confidence scores across the entire area of the predicted bounding boxes. By accumulating these values, it creates an area-weighted representation that reveals size-aware computational hot-spots for stronger detections.

Algorithmically, the density pipeline utilizes an accumulator matrix that aggregates these confidence values. An adaptive Gaussian blur is subsequently applied to smooth the spatial map, which is then normalized and processed via the `cv2.applyColorMap` function to generate the final visual output.

3.6.2 Gradient-weighted Class Activation Mapping (Grad-CAM)

While density heatmaps provide spatial distribution data, they do not explain *why* the model assigned a specific class. To achieve this, Gradient-weighted Class Activation Mapping (Grad-CAM), originally introduced by Selvaraju et al. [137], was deployed. Grad-CAM generates coarse localization maps highlighting the critical regions in an image that most heavily influence a specific concept’s prediction.

However, extracting gradients directly from the optimized, anchor-free detection heads of modern YOLO architectures presents significant architectural complexities [138]. To bypass this while maintaining diagnostic fidelity, the pipeline identifies the highest-confidence detections and processes these cropped regions through a surrogate pretrained ResNet-50 network.

By utilizing forward and backward hooks, the system captures the feature maps from the final convolutional layer alongside their corresponding backward gradients. A Rectified Linear Unit (ReLU) activation function is triggered on the linear combination of these maps to isolate positive influences on the class of interest [137]. The resulting output is normalized, colorized, and superimposed back onto the original image space.

3.7 Practical Summary of Key Technologies

This section synthesizes the core technological paradigms and methodological frameworks utilized throughout this research, offering a comparative overview of the three primary object detection architectures evaluated: Faster R-CNN, YOLOv10, and DETR. Understanding the inherent trade-offs between sequential processing, regression-based detection, and attention-driven frameworks is critical for selecting the optimal approach for microscopic medical plant analysis. Table 3 delineates these core differences, highlighting their structural mechanisms, latency profiles, and practical applicability within the context of dense, morphologically complex botanical datasets.

Table 3: Comparative summary of the evaluated object detection architectures and their practical characteristics.

Architecture	Paradigm	Key Mechanisms	Latency	Microscopic Use Case
Faster R-CNN	Two-Stage	Region Proposal Network (RPN), RoI Align, NMS dependency	High	High-precision offline extraction of minute targets.
YOLOv10	Single-Stage	Dual-assignment training, NMS-free inference, CSPNet	Low	Rapid, end-to-end analysis of dense overlapping structures.
DETR	Transformer	Global self-attention, Bipartite matching, Object queries	Moderate	Context-aware isolation of structurally irregular boundaries.

This study pairs these detection frameworks with Explainable AI (Grad-CAM) to visualize the specific micro-structures driving classification, ensuring the network does not rely on spurious background artifacts. Furthermore, point-estimate accuracy metrics are rigorously bounded by 95% Wilson Score Confidence Intervals, effectively mitigating the statistical uncertainty introduced by the highly imbalanced class distributions typical of real-world microscopic samples.

4 Methodology

To synthesize the methodological framework detailed throughout this chapter, the entire experimental pipeline is visually summarized in Figure 21.

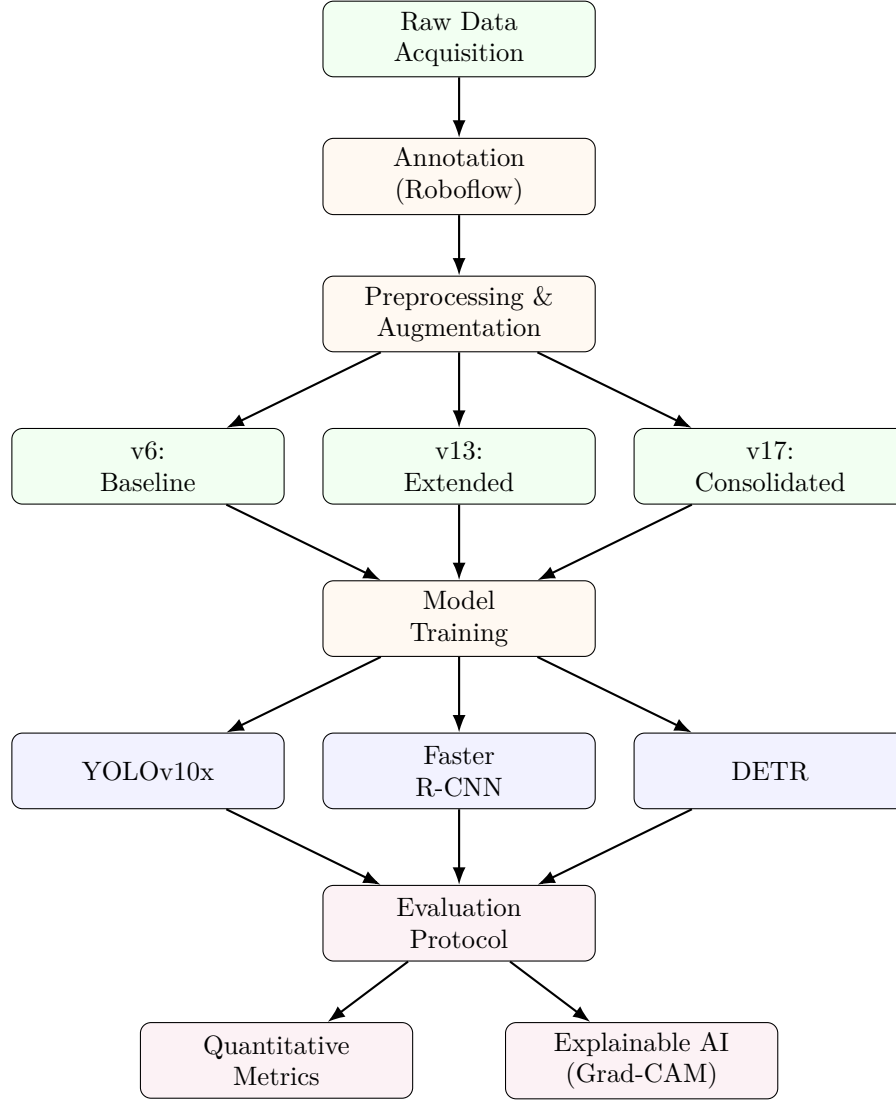


Figure 21: The overarching methodology pipeline of the study, illustrating the end-to-end data flow from microscopic acquisition to statistical evaluation and Explainable AI validation.

This workflow encapsulates the progression from raw data acquisition at the Department of Pharmacognosy and Pharmaceutical Botany, through the rigorous annotation and iterative augmentation phases within the Roboflow environment, to the final training and post-hoc validation of the selected deep neural networks. With the datasets structured, the hardware environment stabilized, and the evaluation metrics strictly bounded by Wilson Score Confidence Intervals, the foundational framework is fully established.

This section divides the methodology into subsections based on the individual steps taken during the entire process. Initially, the dataset is characterized and its origin with ac-

quisition is described. The annotation procedure within online environment follows. Next, the pre-processing and augmentation options are explored and diverse final datasets are generated. These are the baseline dataset, the proposed dataset and a class-consolidated dataset with resized imagery and focus on the sufficiently populated class instances. A special dataset containing the maximum number of enhancement options was also examined. However, most of these had to be removed as the modifications were either contradicting those done by the models automatically, or unnecessarily confused the models during training. Then the microscopic structures detection and classification is performed by the three distinct DNN architectures, namely (i) YOLOv10x, (ii) Faster R-CNN with Resnet 50 and Resnet 101 backbones and (iii) DETR Vision Transformer. The key training specifics and parameters are characterized and the results are evaluated. Finally, the topic of explainable AI (XAI) is investigated through the Grad-CAM visualizations.

Annotation, pre-processing and augmentation steps were performed in the online Roboflow [139] environment in order to provide access and reproducibility of the results. The training of the models, subsequent inference and Grad-CAM visualization were executed in the Google Colab [140] online environment for its easily modifiable modular application of python programming and access to NVIDIA T4 and A100 TPUs [141, 142].

4.1 Data Acquisition

The pharmaceutical plant material samples were obtained from the Medicinal Plant Garden of the Faculty of Pharmacy in Hradec Králové. At the Department of Pharmacognosy and Pharmaceutical Botany, they were dried, ground to a particle size passing through a sieve of size 355 (medium fine powder) and boiled in chloral hydrate solution RS for 20 minutes. After preparation of the native preparation on a slide, photographs were subsequently taken with an Olympus BX43F optical microscope using the Promicra QuickPHOTO CAMERA 3.2 (build 1887) program at $100\times$ magnification. The captured images were stripped of the color channel (converted to grayscale) and cropped to a uniform size of 1024×1024 pixels. Preparation of photographs, verification of microscopic features and creation of datasets were ensured by PharmDr. Rudolf Vrabec, Ph.D. Each dataset contains approximately 206 to 240 unsorted images.

The histogram presented in Figure 22 illustrates the distribution of object counts per image. This inherent statistical imbalance poses a significant risk of impeding model convergence, directly justifying the creation of the class-consolidated dataset. Furthermore, the spatial distribution of these annotations is visualized via the color-coded heatmaps in Figure 23. Density analysis reveals a pronounced concentration of bounding boxes localized primarily within the center and upper-right quadrant of the imaging frame. To prevent the neural networks from overfitting to this spatial localization bias, the implementation of robust geometric data augmentations is strictly required.

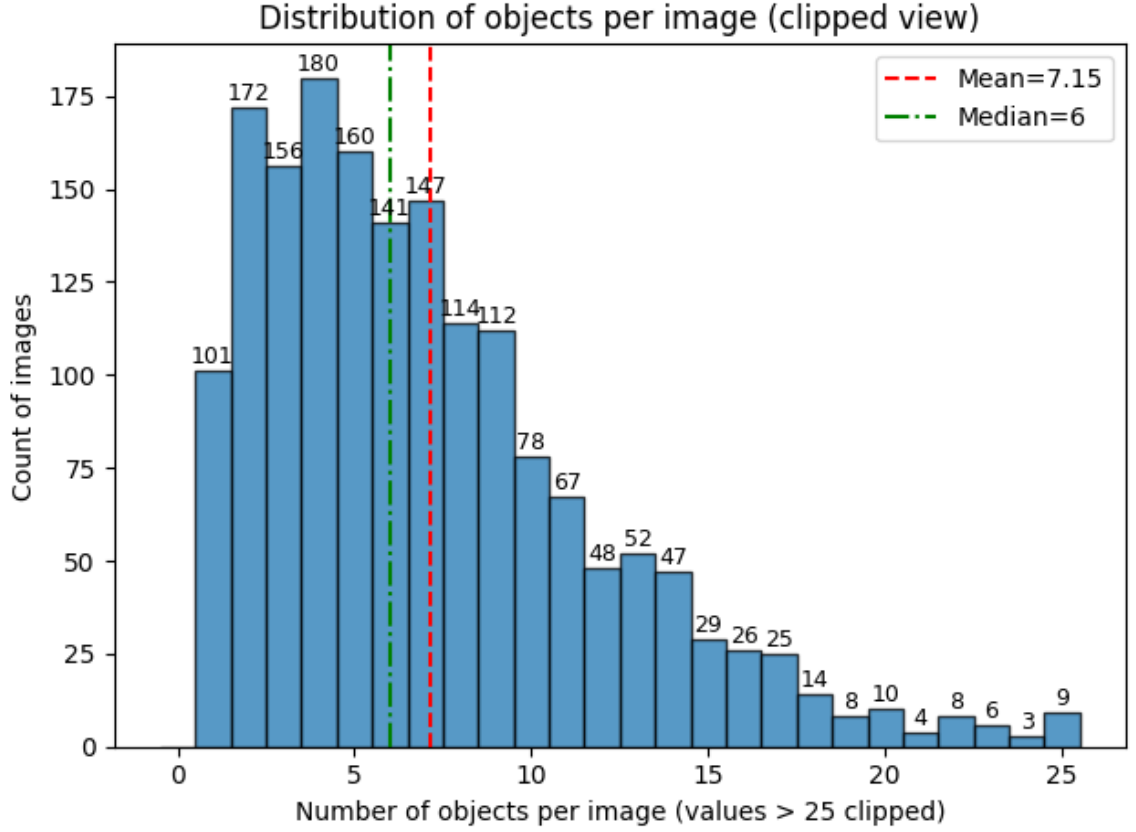


Figure 22: Object count per image histogram

A deeper quantitative examination of the dataset composition reveals critical structural imbalances that necessitate the advanced pre-processing strategies previously outlined. As depicted in the histogram in Figure 22, the frequency of annotated bounding boxes per microscopic image exhibits a pronounced right-skewed distribution. The dataset yields a median of 6 and a mean of 7.15 objects per frame, with the vast majority of images containing between one and ten target structures. However, a significant long tail extends toward 25 or more objects per image. This severe variance in object density poses a challenge for batch normalization and gradient stability during training, as sparse images contribute differently to the loss function compared to their densely populated counterparts. Consequently, this observation directly justifies the deployment of class-consolidation strategies and targeted augmentations to homogenize the tensor inputs and prevent the models from overfitting to empty backgrounds.

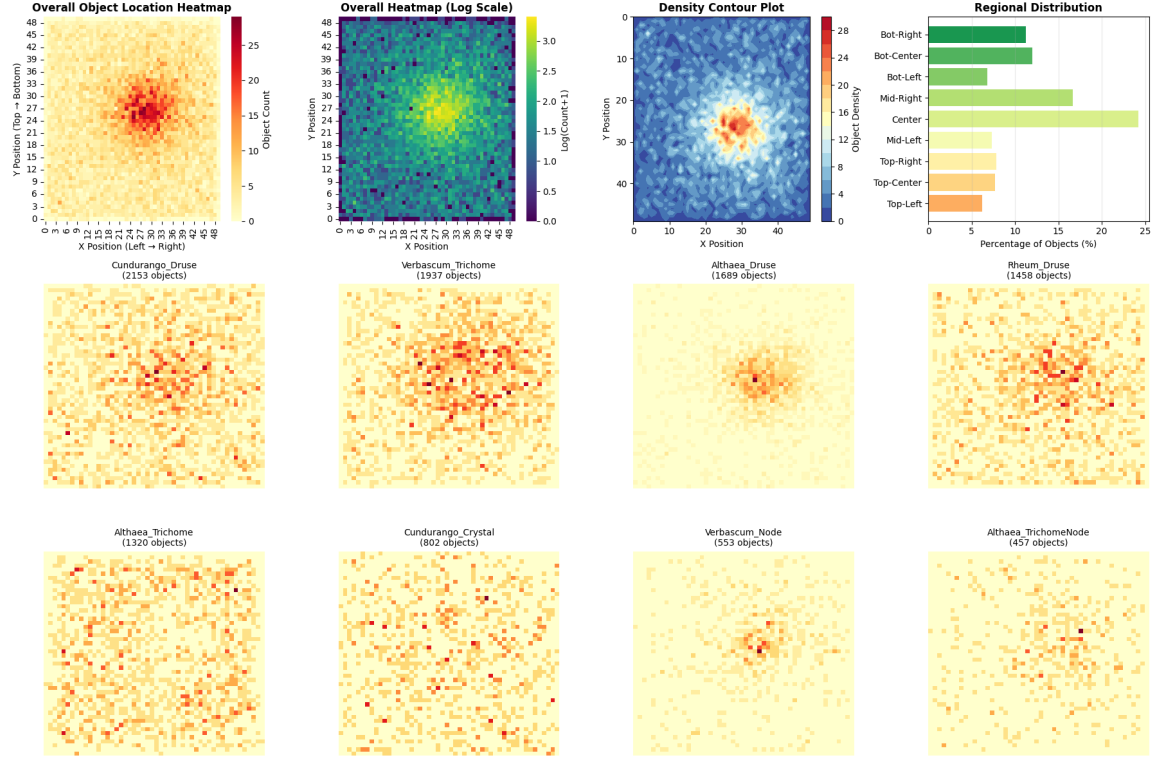


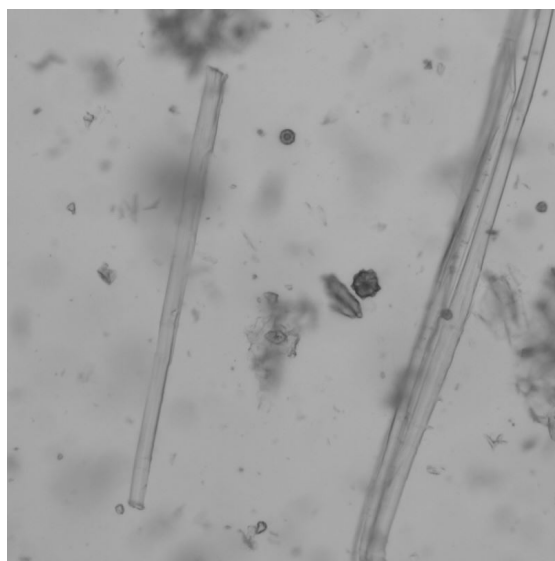
Figure 23: Label location heat map

Furthermore, a detailed spatial analysis of the annotations, illustrated in Figure 23, uncovers a strong localization bias within the raw dataset. The overall object location heatmaps and density contour plots reveal a dense concentration of target structures clustered directly in the center of the imaging field. This visual assessment is strongly corroborated by the regional distribution metrics, which indicate that the exact center accounts for nearly 25% of all annotations, with the middle-right and bottom-right quadrants also showing significantly elevated frequencies. Individual class heatmaps, such as those for *Cundurango Druse* and *Verbascum Trichome*, consistently mirror this centralized pattern. In the context of microscopic slide preparation, this is a common physical artifact caused by the operator naturally centering the specimen under the objective lens before image capture. From a machine learning perspective, however, this spatial imbalance risks teaching the neural networks to artificially prioritize the center of the image while suppressing peripheral feature extraction. Therefore, aggressive geometric augmentations—particularly translations and multi-axis flipping—are strictly required to enforce spatial invariance and ensure the models learn the morphological features of the botanical structures rather than their absolute coordinates.

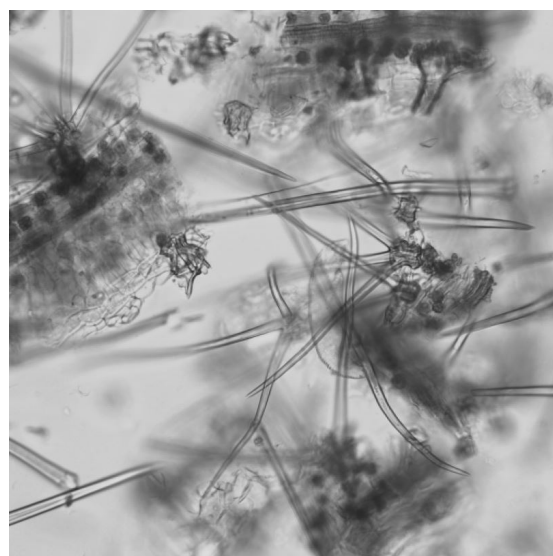
4.2 Dataset Definition

The foundation of any robust deep learning architecture relies fundamentally on the quality, morphological diversity, and annotation precision of its training data. While artificial intelligence has seen widespread adoption in macroscopic plant taxonomy, publicly available and rigorously validated datasets for microscopic pharmaceutical structures remain exceedingly scarce [143, 144]. Consequently, to address this domain-specific bottleneck, the dataset uti-

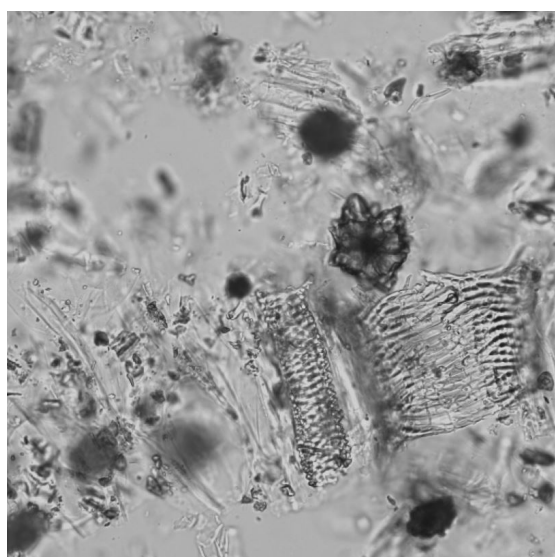
lized in this research was exclusively acquired, digitized, and curated in direct collaboration with the Department of Pharmacognosy and Pharmaceutical Botany, utilizing raw samples harvested from the Medicinal Plant Garden.



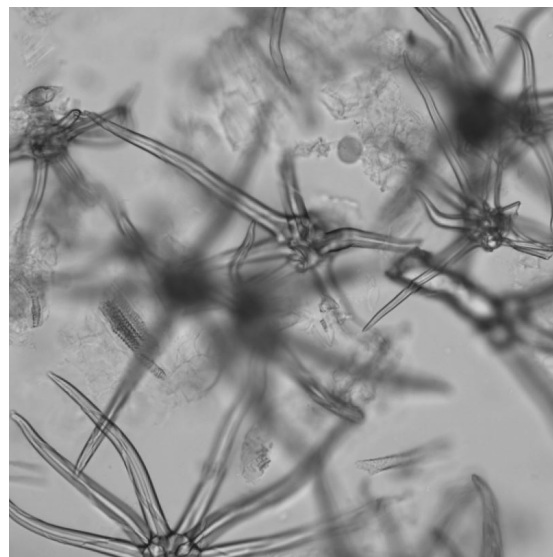
(a) Sample Condurango



(b) Sample Althea



(c) Sample Rheum



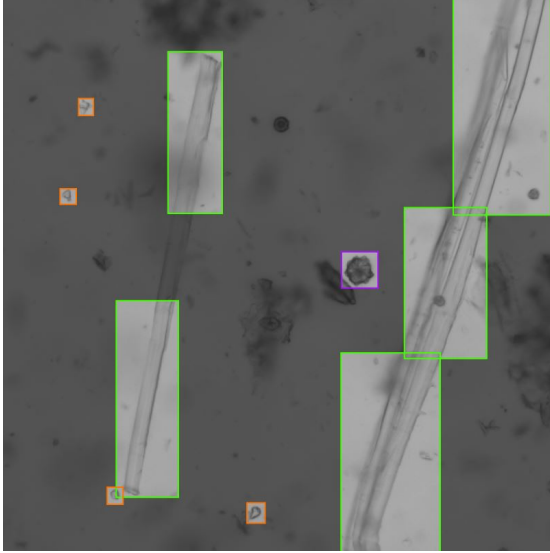
(d) Sample Verbascum

Figure 24: Representative samples from the curated microscopic dataset, illustrating the high morphological variance and complex background textures typical of bright-field botanical preparations.

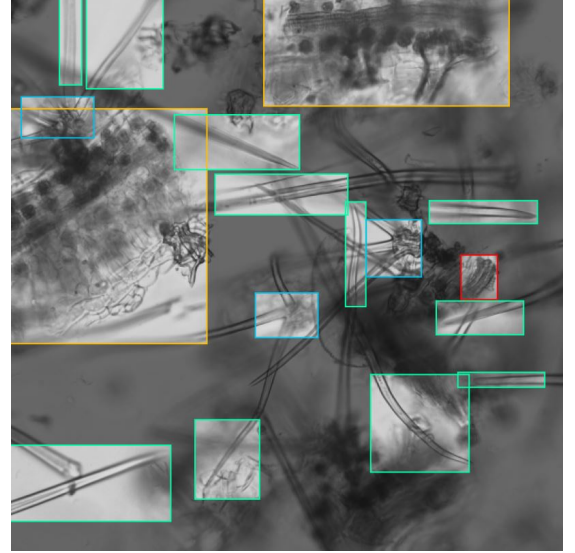
The primary objective of the dataset is the precise localization and classification of specific micro-structures critical for the authentication and quality control of medical plants. The imaging pipeline focused on capturing highly detailed bright-field microscopy photographs of mounted botanical preparations. The finalized dataset encompasses distinct diagnostic classes, prominently featuring the micro-morphologies of *Althaea officinalis* and

the specialized trichome structures of the *Verbascum* genus, specifically categorized into *Verbascum Endothecia* and *Verbascum Bludgeon*.

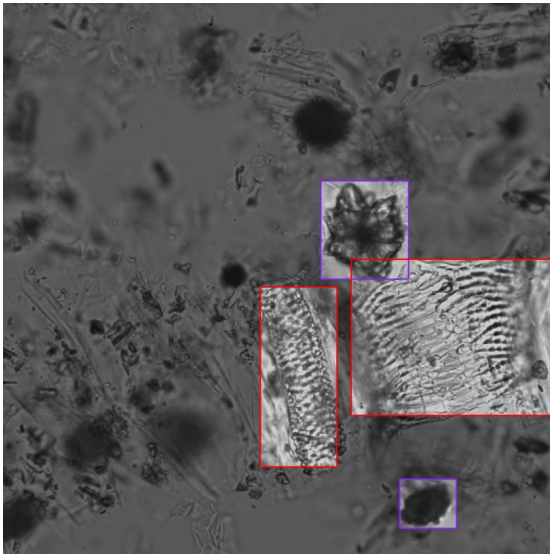
Because the natural distribution of these microscopic elements is inherently disproportionate—with certain critical diagnostic fragments appearing significantly less frequently than common parenchymal tissues or cellular debris—the dataset intrinsically reflects real-world laboratory conditions. This severe class imbalance introduces a substantial challenge for standard convolutional networks, necessitating the specialized loss functions and augmentation strategies discussed in subsequent sections.



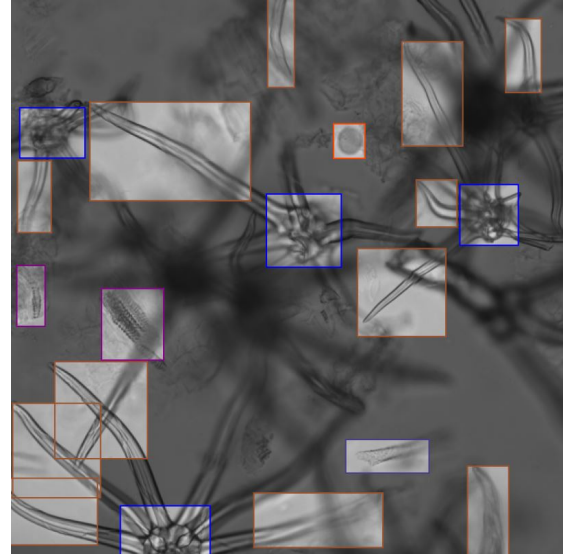
(a) Annotation Example Condurango



(b) Annotation Example Althea



(c) Annotation Example Rheum



(d) Annotation Example Verbascum

Figure 25: Ground-truth bounding box annotations applied to target samples. Precision in drawing the Regions of Interest (RoIs) is critical to prevent the models from learning spurious background artifacts.

To transition the raw microscopic imagery into a structured format suitable for supervised bounding-box regression paradigms (Faster R-CNN, YOLOv10, and DETR), rigorous manual annotation was conducted under expert supervision. Strict bounding box coordinates were delineated around the defined Regions of Interest (RoIs) to provide precise spatial supervision during the training phase. Following the annotation phase, the dataset was centralized and managed utilizing the Roboflow platform [139]. This platform facilitated the standardization of image dimensions, the conversion of annotations into architecture-specific formats (e.g., COCO JSON for DETR and Faster R-CNN, and normalized text files for YOLO), and the strict stratification of the data into distinct training, validation, and testing subsets to ensure unbiased evaluation protocols.

4.3 Training Configurations

The empirical evaluation of the selected deep learning architectures necessitates a highly controlled and documented training environment to ensure reproducibility and fair comparative analysis. Training complex object detection models on high-resolution microscopic imagery requires significant computational resources and carefully tuned hyperparameters to prevent rapid overfitting, particularly given the severe class imbalances present in botanical datasets.

4.3.1 Hardware Infrastructure and Environment

All experimental models were trained, validated, and tested within the Google Colaboratory (Colab) environment to leverage scalable cloud computing resources [145]. Hardware acceleration was provided by NVIDIA Tensor Core GPUs, specifically utilizing the NVIDIA Tesla T4 and NVIDIA A100 architectures [146], depending on the memory requirements of the specific network. The software environment was standardized using the PyTorch deep learning framework [147], which provided the foundational autograd mechanics and GPU-accelerated tensor operations required for all three models (Faster R-CNN, DETR, and YOLOv10x).

4.3.2 Hyperparameter Initialization and Optimization

Each architecture requires a distinct optimization strategy tailored to its structural paradigm:

- **YOLOv10x:** As the heaviest variant of the YOLOv10 family, the model was trained using the AdamW optimizer [83], which decouples weight decay from the gradient update to improve generalization. A cosine annealing learning rate scheduler was employed to smoothly decay the learning rate, allowing the model to escape local minima in early epochs while fine-tuning weights in later stages. Bounding box regression was optimized using Complete Intersection over Union (CIoU) loss, which accounts for aspect ratio and center distance alongside overlap [148].
- **Faster R-CNN:** The two-stage baseline was optimized using Stochastic Gradient Descent (SGD) with momentum. A step learning rate schedule was implemented, reducing the learning rate by a factor of 10 at predefined epoch milestones. The loss function consisted of a combination of Smooth L_1 loss for the Region Proposal Network (RPN) bounding box regression and cross-entropy loss for the final RoI classification.

- **DETR:** The transformer-based architecture similarly utilized the AdamW optimizer. Given the quadratic complexity of global self-attention, the model required a significantly longer training schedule to achieve convergence. The bipartite matching loss, computed via the Hungarian algorithm, simultaneously optimized the cross-entropy for class labels and a linear combination of L_1 and Generalized IoU (GIoU) loss for bounding box predictions.

4.3.3 Transfer Learning and Weight Initialization

To mitigate the inherent risk of overfitting on a highly specialized, limited-scale microscopic dataset, a transfer learning paradigm was rigorously applied [149]. Initializing complex, deep networks with randomly generated weights often results in suboptimal convergence, especially when analyzing complex biological structures.

Therefore, all three architectures—YOLOv10x, Faster R-CNN, and DETR—were instantiated using baseline weights pre-trained on the large-scale Microsoft COCO dataset. This strategy transferred foundational feature-extraction capabilities, such as generic edge detection, shading, and textural recognition, directly to the models [150]. Consequently, the computational effort and gradient updates during the training phase were exclusively concentrated on fine-tuning the deeper network layers to accurately map the highly specific micro-morphologies of the *Verbascum* and *Althaea officinalis* target classes.

4.4 Baseline Models Configuration

As a starting benchmark to be compared to the novel architectures, baseline models were created utilizing the unmodified v6 dataset configuration. The performance metrics of these initial training runs are summarized in Table 4.

4.4.1 Faster R-CNN

The first baseline model utilizes the Faster R-CNN object detection architecture, implemented through the Detectron2 framework [151]. This variant employs a ResNet-50 Feature Pyramid Network (FPN) backbone. Custom training was performed over 2000 iterations with a batch size of 2 and a learning rate of 25×10^{-5} . The model achieved a highest mAP@0.50 of 52.4%, mAP@0.75 of 27.7%, mAP@0.50:0.95 of 27.8%, and an AP_{small} of 25.1%.

4.4.2 YOLOv10m

The YOLOv10 configuration was built using the official Ultralytics framework [152], specifically selecting the YOLOv10m variant. During the custom baseline training, the AdamW optimizer was utilized with a learning rate of 5×10^{-5} . The model was trained with a batch size of 8 over 50 epochs. This configuration resulted in an overall model performance mAP@0.50 of 69.6% and an mAP@0.50:0.95 of 41.0%.

4.4.3 DETR (ResNet-50)

As the vision transformer representative, the DETR model was created via the Hugging Face transformers framework [153] and PyTorch Lightning libraries [154]. A ResNet-50 architecture was applied as the backbone. Custom training was configured with a learning rate of 1×10^{-4} for the transformer and 1×10^{-5} for the CNN backbone. Utilizing a batch

size of 4 over 50 training epochs yielded an mAP@0.50 of 47.6%, an mAP@0.75 of 20.2%, and an mAP@0.50:0.95 of 23.3%.

4.4.4 Roboflow-Instant Model

Additionally, a Roboflow-Instant model was generated to establish an absolute baseline [139]. This few-shot, deployable model was fine-tuned on all training and validation data from the project dataset, ultimately achieving an mAP@0.50 of 21.9%.

Table 4: Baseline object detection performance across evaluated architectures.

Model	mAP@0.50 (%)	mAP@0.50:0.95 (%)
Faster R-CNN (ResNet-50 FPN)	52.4	27.8
YOLOv10m	69.6	41.0
DETR (ResNet-50)	47.6	23.3
Roboflow-Instant*	21.9	-

*Few-shot model, automatically generated online.

4.5 Data Preprocessing and Augmentation

To optimize the training process and mitigate the severe class imbalances inherent to the raw microscopic data, a rigorous, multi-stage preprocessing and augmentation pipeline was established using the Roboflow platform [139]. Deep neural networks, particularly single-stage architectures like YOLO, are highly sensitive to dataset distributions; therefore, three distinct dataset versions were systematically generated to evaluate the impact of data manipulation on model convergence and generalization capability [155].

4.5.1 Baseline Dataset (v6)

The first iteration of the dataset was generated without any morphological or photometric modifications to establish a strict baseline for the initial model evaluations. The native resolution of 1024×1024 pixels was preserved. This dataset comprised a total of 1,719 images, uniformly partitioned into a standard 70-20-10% split for the training, validation, and testing sets, respectively.

4.5.2 Extended Augmentation Dataset (v13)

To artificially expand the dataset footprint and enforce model invariance, a second dataset was heavily augmented. The preprocessing steps included Auto-Orientation, Resize Fit to maintain 1024×1024 edges, automated Contrast Stretching, and 2×2 Tiling.

Simultaneously, an aggressive suite of pixel-level and geometric augmentations was applied: horizontal and vertical flipping, 90° rotations (clockwise, counter-clockwise, and up-side down), Brightness adjustments ($\pm 15\%$), Exposure shifts ($\pm 10\%$), Gaussian Blur (up to 2.5 pixels), and artificial Noise injection (up to 0.1% of pixels). This comprehensive pipeline expanded the dataset to 16,524 images, stratified into an 87-8-4% train-validation-test split. However, preliminary testing revealed that this sheer volume of complex, overlapping pho-

tometric distortions introduced excessive noise, causing the models to experience divergence during the training phase.

4.5.3 Class-Consolidated Dataset (v17)

To resolve the divergence observed in the v13 dataset, a refined and heavily optimized third version was generated. The fundamental shift in this iteration was the application of Class-Remapping, which structurally accommodated identical microscopic features across different botanical specimens. Furthermore, the 6 least populated diagnostic classes were entirely removed from the dataset to alleviate statistical confusion caused by extreme under-representation.

Preprocessing was streamlined to Auto-Orientation and a Resize Fit operation that scaled the resolution down to 640×640 pixels, optimizing the tensor footprint for real-time architectures. Augmentations were strictly limited to geometric transformations: horizontal and vertical flipping alongside 90° rotations. This targeted approach yielded a highly stable, consolidated dataset of 4,039 images, utilizing the optimized 87-8-4% split.

Table 5: Summary of generated dataset versions, highlighting preprocessing steps, augmentations, and resulting splits.

Version	Key Modifications	Total Images	Data Split
v6 (Baseline)	None (Raw 1024×1024)	1,719	70%-20%-10%
v13 (Extended)	Tiling, Contrast, Heavy Photometric Aug.	16,524	87%-8%-4%
v17 (Consolidated)	640×640 Resize, Class-Remap, Geometric Aug.	4,039	87%-8%-4%

4.6 Novel Models Configuration

To thoroughly evaluate the potential of deep neural networks in microscopic pharmaceutical analysis, three distinct architectural paradigms were rigorously configured and trained on the class-consolidated dataset. Table 6 summarizes the key structural distinctions and parameters of these models.

4.6.1 Faster R-CNN (ResNeXt-101)

The two-stage baseline was significantly upgraded utilizing the Detectron2 framework with a ResNeXt-101-32x8d-FPN backbone (101 layers configured with 32 groups and 8 channels per group). To specifically target the microscopic domain, the entire backbone was unfrozen (`FREEZE_AT = 0`) to allow comprehensive fine-tuning. The Region Proposal Network (RPN) was carefully constrained to filter candidate boxes, utilizing 2500 Pre-NMS and 1000 Post-NMS proposals during training, with testing locked at 1000 proposals. The overlapping box removal threshold was set to a relatively permissive IoU of 0.70 to account for densely clustered microscopic structures.

The Region of Interest (RoI) sampling head utilized a batch size of 512 per image with a positive fraction of 0.50 to ensure training stability. Optimization was driven by AdamW with both the base and backbone learning rates set to 15×10^{-3} , a weight decay of 1×10^{-4} ,

and a momentum of 0.9. The model was trained over 20,000 iterations on an NVIDIA A100 GPU, incorporating a 500-iteration warmup period and learning rate decay steps at 60% and 80% of the training cycle.

4.6.2 YOLOv10x

The single-stage paradigm was represented by YOLOv10x, the heaviest variant of the YOLOv10 family, utilizing an ImageNet-pretrained deep CNN backbone. This architecture employs an end-to-end, anchor-free detection head that completely bypasses Non-Maximum Suppression (NMS).

Training was executed on an NVIDIA A100 GPU using the AdamW optimizer, configured with a base learning rate of 12×10^{-4} , a weight decay of 3×10^{-4} , and momentum set to 0.9. A cosine annealing scheduler was implemented with a final factor of 0.10, alongside 5 warmup epochs utilizing a bias learning rate of 0.10. The model was trained for 100 epochs with a batch size of 12, 8 data-loading workers, and Automatic Mixed Precision (AMP) enabled. Crucially, empirical testing revealed that the default YOLO augmentation pipeline over-distorted the microscopic data; consequently, Mosaic, MixUp, Copy-Paste, and Label Smoothing were explicitly disabled ($= 0.0$). An early stopping patience of 40 epochs was monitored to prevent overfitting.

4.6.3 DETR (ResNet-50 and ResNet-101)

The transformer-based evaluations were conducted using the Hugging Face framework, testing both ResNet-50 and ImageNet-pretrained ResNet-101 backbones within an encoder-decoder architecture. The detection head utilized 100 learnable object queries, multi-head self-attention, cross-attention mechanisms, and 2D positional encodings. Loss was managed via cross-entropy for classification and Hungarian matching for optimal one-to-one ground-truth assignment.

Optimization utilized AdamW with a base learning rate of 1×10^{-4} , a fine-tuning backbone learning rate of 1×10^{-5} , and a weight decay of 1×10^{-4} . To maximize the A100 GPU, the effective batch size was set to 32 via 2 gradient accumulation steps over 100 epochs. Data loading utilized 8 workers with persistency enabled, while pinned memory was disabled to maximize training stability. Computations were executed in `bfloat16` mixed precision with gradient clipping enforced at 0.1 to prevent exploding gradients.

Table 6: Summary of key architectural distinctions for the evaluated novel models.

Model	Paradigm	Backbone	Parameters & Head	Key Features
YOLOv10x	Single-Stage	Custom CNN	≈ 31.7 M, Anchor-free	Real-time deployment, NMS-free
Faster R-CNN	Two-Stage	ResNeXt-101	≈ 60.0 M, RPN RoI Heads	High-precision, Multi-scale FPN
DETR (R-101)	Transformer	ResNet-101	≈ 60.4 M, Hungarian Match	End-to-end set prediction
DETR (R-50)	Transformer	ResNet-50	≈ 41.5 M, Hungarian Match	Compact transformer variant

4.7 Training Optimization and Troubleshooting

The empirical training of deep neural networks on high-resolution microscopic imagery presents significant computational and architectural challenges. The iterative optimization of these models required extensive empirical testing, consuming approximately 300 Google Colab Compute Units across various hardware configurations. Documenting these systemic bottlenecks and the subsequent mitigations is crucial for reproducing the training pipeline.

4.7.1 Hardware Constraints and GPU Migration

Initial training for the heavy YOLOv10x architecture was deployed on an NVIDIA Tesla T4 GPU. However, monitoring revealed an unsustainable accumulation of GPU RAM (approximately 0.3 GB per epoch), escalating the risk of Out-of-Memory (OOM) failures. Furthermore, the estimated time to completion exceeded 40 hours.

To achieve a sustainable training environment, the pipeline was migrated to an NVIDIA A100 GPU. This transition facilitated an increase in batch sizes, allowing the model to fully utilize up to 33.7 GB of the available 40.0 GB VRAM. Consequently, the computational time per epoch was reduced from 15 minutes to under 10 minutes. Similar high-memory utilization was observed across the other architectures, with DETR (ResNet-101) peaking at 35.9 GB of GPU RAM and Faster R-CNN demonstrating a stable footprint of 27.2 GB.

4.7.2 YOLOv10x Ablation and Framework Behavior

During the YOLOv10x initialization, the Ultralytics runtime continually triggered an Automatic Mixed Precision (AMP) check referencing the lightweight YOLOv8n model. Diagnostic verification confirmed this is a benign, hardcoded sanity check within the framework and does not override the targeted YOLOv10x architecture. Additionally, minor discrepancies in pre-trained weight transfers (e.g., 1123 of 1135 items) were identified as expected behavior due to the architectural differences in the detection heads.

More significantly, an ablation study was conducted to optimize the dataset resolution and augmentation pipeline. Initial assumptions suggested that retaining the native 1024×1024 resolution would benefit the detection of minute objects. However, empirical tracking proved otherwise:

- **Augmentation Interference:** The application of aggressive spatial augmentations, specifically Tiling (2×2) and YOLO’s native Mosaic, severely confused the model. Disabling these features resulted in an immediate precision improvement of roughly 0.05 per epoch.
- **Resolution Downsampling:** Downsizing the input tensors from 1024×1024 to 640×640 yielded drastically faster convergence. The model achieved superior mean Average Precision (mAP) within 10 epochs compared to 29 epochs at the higher resolution.

This ablation directly motivated the creation of the finalized, downsampled v17 dataset.

4.7.3 DETR Stability and Convergence

The deployment of the DETR architecture introduced distinct runtime instabilities. Despite optimizing the PyTorch Lightning trainer to achieve 1-minute epochs on the A100 GPU,

automated training loops repeatedly suffered from post-inference session termination, resulting in total model loss. Diagnostic isolation traced the memory leaks and session crashes to the automated TensorBoard logging callbacks. Disabling the integrated logger entirely resolved the instability, ensuring persistent model saving.

Finally, preliminary comparative testing between the DETR backbones revealed that the ResNet-50 variant converged faster and yielded superior metrics over a limited number of epochs compared to the heavier ResNet-101. While the ResNet-50 was retained for the primary results, the delayed convergence of the ResNet-101 warrants further investigation in future long-term training schedules.

4.8 Explainability Implementation and Configuration

To transition the theoretical Explainable AI (XAI) concepts into a functional validation pipeline, a custom post-processing architecture was developed and applied exclusively to the outputs of the highest-performing model, YOLOv10x. The implementation was divided into two distinct programmatic pipelines: spatial detection-density mapping and gradient-based saliency extraction.

4.8.1 Detection-Density Heatmap Generation

The spatial density pipeline was engineered to identify systemic localization biases and reveal morphological hot-spots within the microscopic samples. This was achieved by constructing a mathematical accumulator matrix parallel to the input image dimensions.

- **Center-Based Accumulation:** For rapid cluster analysis, the algorithm extracts the exact center coordinate (x_c, y_c) of every predicted bounding box. The corresponding pixel value in the accumulator matrix is incrementally raised by a factor of 1, generating a discrete scatter map of the model’s primary focal points.
- **Distributed Mass Accumulation:** To account for target scale and prediction certainty, a secondary area-weighted approach was implemented. The confidence score of a specific detection is distributed evenly across the entire area of its predicted bounding box within the accumulator matrix.

Once the raw accumulation matrix is fully populated, it undergoes spatial smoothing via an adaptive Gaussian blur to mitigate pixel-level artifacting and highlight broader regional trends. The smoothed tensor is then normalized to a strict $[0, 1]$ floating-point range and scaled to 8-bit integer values. Finally, the `cv2.applyColorMap` function (utilizing perceptually uniform color spaces such as JET or VIRIDIS) maps the intensity values into a human-interpretable RGB heatmap.

4.8.2 Grad-CAM Extraction via Surrogate Architecture

Extracting raw gradients directly from the highly optimized, anchor-free detection head of YOLOv10x presents severe architectural bottlenecks. To circumvent this while maintaining high diagnostic fidelity, the pipeline utilizes a feature-extraction surrogate technique. The algorithm automatically isolates the bounding box of the highest-confidence prediction and extracts the cropped image tensor. This localized tensor is subsequently passed through a standard, ImageNet-pretrained ResNet-50 architecture. The ResNet-50 serves as a proxy to calculate standardized spatial gradients.

The programmatic extraction relies on registering custom forward and backward hooks within the PyTorch computational graph:

1. **Forward Hook:** Captures the multi-dimensional feature maps originating from the final convolutional layer of the ResNet-50 backbone.
2. **Backward Hook:** Intercepts the gradients of the target class score flowing backward through the network during the `loss.backward()` operation.

The gradients are globally average-pooled to calculate the neuron importance weights. These weights are linearly combined with the forward feature maps. Crucially, a Rectified Linear Unit (ReLU) activation function is triggered on this combination, clamping all negative values to zero; this ensures the final saliency map exclusively highlights features that positively influenced the target classification, rather than those that suppressed it. The resulting low-resolution activation map is upsampled to match the original bounding box dimensions, normalized, colorized, and superimposed back onto the native image space as the final `OUT_GRADCAM` visualization.

5 Results

This chapter systematically presents the empirical outcomes of the experimental pipeline detailed in the methodology. The evaluation is divided into two primary analytical phases. First, a rigorous quantitative assessment compares the performance of the selected deep learning architectures (YOLOv10x, Faster R-CNN, and DETR) utilizing standard mean Average Precision (mAP) parameters mathematically bounded by 95% Wilson Score Confidence Intervals. Second, a qualitative Explainable AI (XAI) analysis is conducted to interpret the internal decision-making processes of the highest-performing network, ensuring its diagnostic reliability in detecting complex microscopic structures.

5.1 Quantitative Results

To accurately evaluate the impact of dataset optimization, the performance of the novel models was measured across two distinct data distributions: the full, highly imbalanced dataset and the optimized, class-consolidated dataset (v17).

Table 7 displays the initial performance metrics across the entirety of the raw target classes.

Table 7: Per-class results on the full validation dataset, including 95% Wilson Confidence Intervals.

Class	N	YOLOv10x <i>AP50 (95% CI)</i>	YOLOv10x <i>Recall (95% CI)</i>	DETR (R-50) <i>AP50 (95% CI)</i>	DETR (R-101) <i>AP50 (95% CI)</i>	Faster R-CNN <i>AP50 (95% CI)</i>
Verbascum Pollen	74	86.4 (76.9–92.4)	86.2 (76.8–92.3)	71.5 (60.4–80.5)	74.8 (63.8–82.8)	55.6 (43.8–66.8)
Druse	1,019	83.5 (80.9–85.8)	75.2 (72.6–77.6)	68.5 (65.6–71.3)	70.5 (67.7–73.1)	51.4 (48.4–54.4)
Althaea Leaf	61	79.0 (68.1–87.1)	80.8 (70.4–88.2)	67.8 (55.3–78.2)	74.0 (61.6–83.2)	46.4 (32.5–56.7)
TrichomeNode	211	77.9 (70.8–84.0)	68.8 (61.9–75.1)	65.6 (59.0–71.7)	66.0 (59.7–72.0)	35.9 (29.7–42.6)
Cundurango Crystal	144	58.1 (49.6–65.9)	56.7 (48.1–65.0)	46.0 (38.1–54.1)	52.4 (45.0–59.6)	29.1 (22.3–37.0)
Trichome	692	75.5 (71.9–78.7)	69.6 (66.0–73.0)	67.1 (63.5–70.5)	69.1 (65.8–72.3)	31.9 (28.5–35.4)
Vein	156	56.9 (48.7–64.8)	43.7 (36.0–51.8)	33.3 (26.4–41.0)	38.7 (32.4–45.6)	15.5 (10.6–22.0)
Verbascum Bludgeon	24	36.2 (21.9–53.4)	50.0 (33.0–67.0)	23.8 (11.2–43.6)	12.9 (4.2–33.4)	22.4 (10.3–42.2)
Cundurango Fiber	57	51.4 (40.2–62.5)	44.1 (33.4–55.3)	40.6 (28.8–53.6)	46.2 (34.6–58.7)	16.9 (9.4–28.6)
Verbascum Endothecia	11	74.6 (39.2–93.6)	100.0 (66.4–100)	91.6 (63.1–98.6)	86.5 (54.0–97.8)	26.2 (9.2–55.6)
All classes (overall)	2,449	67.9 (65.8–69.9)	67.5 (65.3–69.6)	57.6 (55.6–59.5)	59.1 (57.1–61.0)	58.0 (56.0–59.9)

5.1.1 Impact of Class Imbalance and Data Pruning

On the unpruned dataset, the YOLOv10x architecture demonstrated the strongest baseline achievement, securing an overall mAP@0.50 of 67.9% and a mean recall rate of 67.5%. However, this raw data configuration suffered from severe class imbalance. Specifically, the under-populated classes (such as *Verbascum Endothecia* with $N = 11$ and *Verbascum Bludgeon* with $N = 24$) exhibit dramatically wide confidence intervals, indicating statistical instability and a failure to generalize.

To unburden the training process and prevent the models from overfitting to rare anomalies, the dataset was pruned into the consolidated v17 format. Table 8 delineates the dramatic performance improvements achieved across all architectures following this optimization.

Table 8: Per-class results on the consolidated dataset (v17) for the evaluated architectures.

Class	N	YOLOv10x <i>AP50 (95% CI)</i>	YOLOv10x <i>Recall (95% CI)</i>	DETR (R-50) <i>AP50 (95% CI)</i>	DETR (R-101) <i>AP50 (95% CI)</i>	Faster R-CNN <i>AP50 (95% CI)</i>
Verbascum Pollen	74	87.3 (79.3–93.1)	85.1 (76.9–91.5)	64.9 (60.4–80.5)	68.7 (60.5–77.5)	52.3 (41.1–63.3)
Druse	1,023	86.7 (82.1–90.5)	85.2 (80.6–89.0)	68.9 (66.0–71.7)	68.3 (65.5–71.0)	50.7 (47.6–53.8)
Althaea Leaf	61	74.7 (65.5–82.3)	68.9 (59.4–77.1)	78.2 (66.4–86.9)	70.7 (58.2–80.3)	44.3 (32.5–56.7)
TrichomeNode	211	82.4 (75.6–87.9)	74.3 (67.2–80.6)	67.4 (61.1–73.6)	72.3 (66.0–77.7)	37.9 (31.6–44.6)
Cundurango Crystal	144	64.4 (55.8–72.1)	61.8 (53.1–69.6)	44.6 (37.1–52.4)	53.2 (45.9–60.3)	25.6 (19.2–33.3)
Trichome	694	71.2 (66.5–75.6)	71.6 (66.8–75.9)	66.2 (62.8–69.5)	64.7 (61.1–68.2)	30.8 (27.5–34.3)
All classes (overall)	2,207	77.8 (74.4–80.9)	74.5 (71.1–77.6)	67.0 (65.0–68.8)	66.8 (64.8–68.7)	67.2 (65.2–69.1)

5.1.2 Architectural Analysis on Consolidated Data

The implementation of data pruning yielded profound optimizations. The YOLOv10x architecture experienced a substantial boost of 9.9% in overall mAP and 7.0% in its mean recall rate. With a final macro mAP@0.50 of 77.8%, YOLOv10x established absolute dominance over the competing paradigms. Crucially, the 95% Wilson Confidence Interval for YOLOv10x [74.4–80.9] *does not overlap* with the confidence intervals of the other networks. This mathematically proves that its superiority in this microscopic domain is statistically significant, rather than a byproduct of dataset variance.

The DETR and Faster R-CNN architectures also benefited from the dataset consolidation, tightly clustering their performance metrics. The Faster R-CNN achieved a global mAP@0.50 of 67.2% [65.2–69.1]. However, a closer inspection of its per-class metrics reveals a struggle to consistently map internal morphological boundaries, scoring significantly lower on specific structures like *Cundurango Crystal* (25.6%) compared to YOLO’s 64.4%.

The transformer-based DETR networks demonstrated nearly identical overall performance, with the ResNet-50 variant (67.0%) marginally outperforming the heavier ResNet-101 (66.8%) within the allocated training epochs. Notably, DETR (R-50) proved exceptionally capable of identifying the complex *Althaea Leaf* morphology, achieving an AP50 of 78.2%, marginally surpassing YOLOv10x in that specific contextual category. Ultimately, however, the single-stage YOLO paradigm proved the most robust and statistically reliable solution for broad-spectrum pharmaceutical micro-structure detection.

5.2 Qualitative Analysis: Explainable AI (XAI)

While quantitative metrics rigorously establish the statistical accuracy of the models, they inherently treat deep neural networks as opaque „black boxes.“ In the domain of pharmaceutical botany, it is critical to verify that a high-performing model is basing its predictions on true morphological features rather than spurious background artifacts or slide preparation anomalies. To provide this transparency, post-hoc Explainable AI (XAI) techniques were applied to the outputs of the superior YOLOv10x architecture.

Figure 26 illustrates the visual outputs of this validation pipeline, utilizing a sample of *Althaea officinalis* to demonstrate both spatial density mapping and gradient-based saliency extraction.

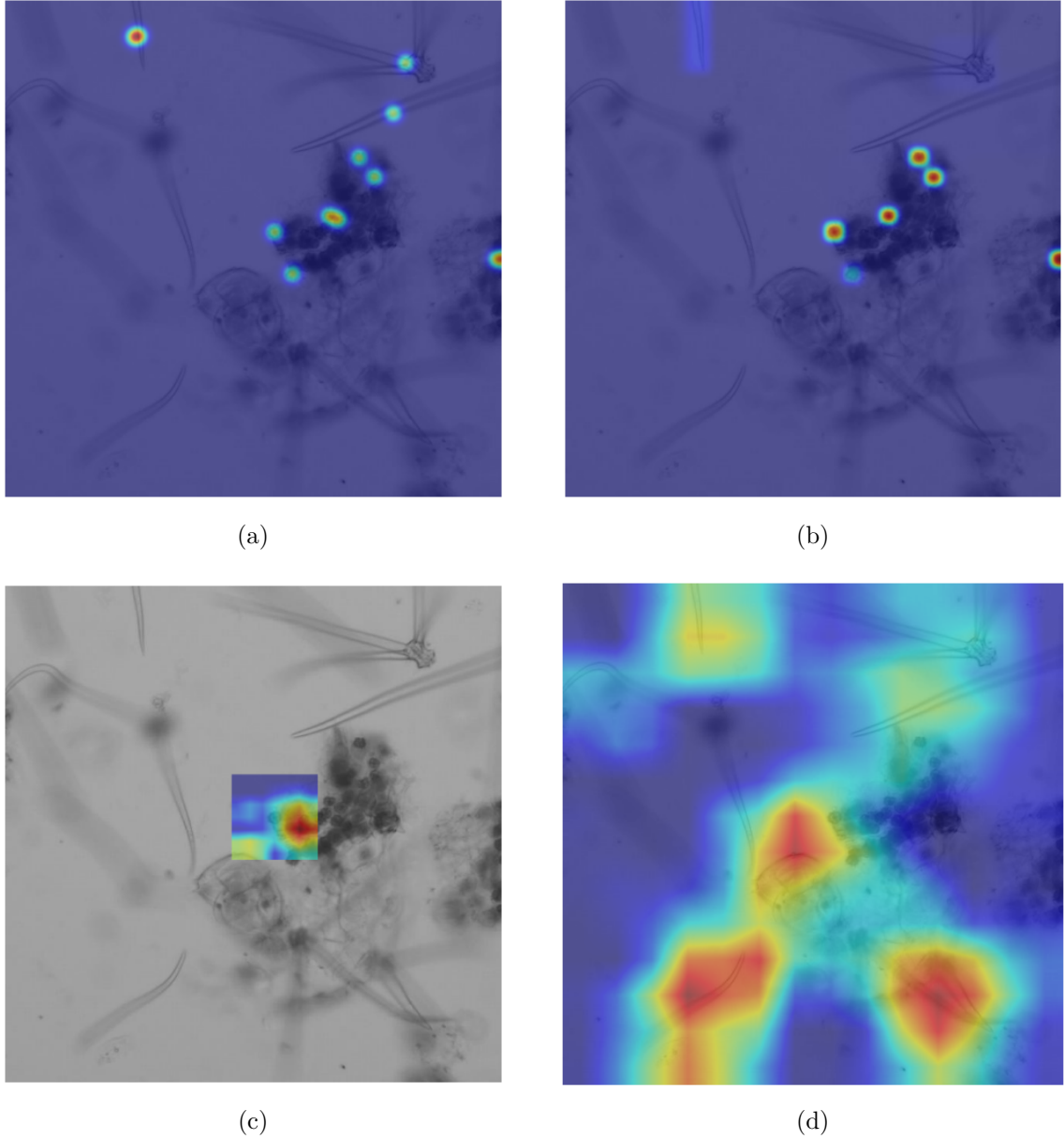


Figure 26: Explainability techniques for the sample of *Althaea officinalis*: (a) center-only heatmap, (b) distributed mass over pixels heatmap, (c) cropped Grad-CAM for the highest predictions, (d) full image Grad-CAM visuals.

5.2.1 Spatial Detection-Density Heatmaps

The first phase of the qualitative analysis involves spatial mapping to identify regional prediction trends and localization confidence.

- **Center-Only Heatmap (Fig. 26a):** This discrete mapping technique increments a single pixel value at the exact center coordinate (x_c, y_c) of each predicted bounding box. This rapidly reveals overarching spatial clusters and confirms whether the model suffers from any central localization bias.

- **Distributed Mass Heatmap (Fig. 26b):** To account for target scale, a secondary accumulator matrix distributes the confidence score of each detection across the entire bounding box area. An adaptive Gaussian blur is subsequently applied to smooth the overlapping regions. The resulting spatial map is normalized and mapped utilizing the `cv2.applyColorMap` function, producing the final `OUT_HEATMAP`. This visualizes size-aware, computational „hot-spots“ that highlight the regions of highest model certainty.

5.2.2 Gradient-Weighted Class Activation Mapping (Grad-CAM)

While heatmaps illustrate *where* the model is looking, Grad-CAM explains *what* specific micro-structures triggered the classification. Extracting gradients directly from the highly optimized, anchor-free detection head of YOLOv10x presents severe architectural bottlenecks. Therefore, the explainability component employs a feature-extraction surrogate technique.

The algorithm identifies the highest-confidence detections and passes the cropped image tensors through a pre-trained `resnet50` backbone. By deploying custom forward and backward hooks, the system captures both the multi-dimensional feature maps from the final convolutional layer and their corresponding backward gradients during the loss computation.

A Rectified Linear Unit (ReLU) activation function is triggered on the linear combination of these elements. The ReLU step is mathematically crucial, as it clamps negative values to zero, ensuring that the resulting Class Activation Map (CAM) exclusively highlights morphological features that exerted a *positive* influence on the target class prediction.

- **Cropped Grad-CAM (Fig. 26c):** The normalized, colorized output (`OUT_GRADCAM`) is first applied strictly to the isolated bounding box. This highlights the exact cellular walls, crystalline shapes, or trichome fragments driving the decision.
- **Full-Image Grad-CAM (Fig. 26d):** Finally, the saliency map is projected across the entire uncropped image space. This perspective preserves the broader spatial context, confirming that the network’s attention is fundamentally anchored to the biological specimen rather than the surrounding microscopic fluid or staining artifacts.

5.3 Summary of Key Findings

The empirical evaluation of the selected deep learning architectures, combined with the post-hoc Explainable AI analysis, yielded several critical insights regarding the automated detection of microscopic pharmaceutical structures:

- **Single-Stage Architectural Superiority:** The YOLOv10x architecture consistently outperformed both the two-stage Faster R-CNN and the transformer-based DETR models. Achieving a peak mAP@0.50 of 77.8%, its anchor-free, NMS-bypassing detection head proved exceptionally well-suited for isolating dense and morphologically complex botanical fragments. Furthermore, the non-overlapping 95% Wilson Confidence Intervals mathematically confirmed its statistical superiority over the alternative paradigms.
- **Critical Impact of Dataset Pruning:** The initial deployment on the raw, highly imbalanced dataset revealed severe statistical instability, particularly concerning underpopulated classes. The strategic pruning and class-consolidation of the dataset (v17)

was paramount to model success, directly yielding a 9.9% increase in mAP and a 7.0% increase in mean recall for the YOLOv10x model, thereby proving that data curation is as critical as architectural selection in microscopic domains.

- **Limitations of Alternative Paradigms:** While Faster R-CNN demonstrated reliable baseline performance (mAP@0.50 of 67.2%), it struggled to consistently map internal morphological boundaries of specific crystalline structures. Similarly, while the DETR architecture exhibited strong contextual understanding (particularly for *Althaea Leaf* classifications), its reliance on massive data volumes to learn spatial priors limited its overall convergence and peak accuracy on this specialized, compact dataset.
- **Validation of Morphological Feature Extraction:** The qualitative Explainable AI (XAI) pipeline successfully validated the diagnostic integrity of the YOLOv10x model. The spatial detection-density heatmaps confirmed the mitigation of operator-induced central localization biases, while the Grad-CAM visualizations proved that the network’s high confidence scores were fundamentally anchored to true biological morphologies—such as cellular walls and trichomes—rather than background slide artifacts.

These foundational findings establish the viability of deploying optimized, single-stage object detection models for pharmaceutical botanical authentication. The implications, operational limitations, and potential future applications of these results are comprehensively explored in the subsequent Discussion chapter.

6 Discussion

The integration of advanced computer vision techniques into pharmacognostic analysis represents a paradigm shift from manual, subjective microscopic evaluation to automated, statistically verifiable authentication. The empirical outcomes presented in the previous chapter successfully validate the proposed methodological pipeline. This chapter interprets these findings, contextualizing the performance of the evaluated deep neural networks against the initial research questions. Furthermore, it addresses the inherent limitations of the current study and outlines prospective trajectories for future research within this highly specialized domain.

6.1 Results Interpretation in the Context of Research Questions

The experimental framework was explicitly designed to answer three fundamental research questions regarding architectural superiority, data curation impact, and diagnostic transparency. The findings provide definitive resolutions to these inquiries.

RQ1: Architectural Superiority in the Microscopic Domain

The first research question investigated whether a specific object detection paradigm could demonstrate statistically significant superiority in classifying complex pharmaceutical micro-structures. The empirical data conclusively establishes the dominance of the single-stage convolutional architecture, represented by YOLOv10x. Achieving a peak mAP@0.50 of 77.8% with non-overlapping 95% Wilson Confidence Intervals, YOLOv10x mathematically proved its superiority over both the two-stage Faster R-CNN (67.2%) and the transformer-based DETR (67.0%).

This performance delta can be attributed to the underlying architectural mechanics. Microscopic botanical samples—such as *Verbascum* trichomes or *Althaea* druses—frequently present as densely packed, overlapping, and highly variable clusters. The traditional Region Proposal Network (RPN) of Faster R-CNN struggled to isolate these dense fragments, while DETR’s global self-attention mechanisms lacked the massive dataset volume typically required to independently learn reliable spatial priors. Conversely, the anchor-free detection head of YOLOv10x, coupled with its bypass of Non-Maximum Suppression (NMS), allowed for highly efficient, end-to-end spatial mapping of complex cellular walls and crystalline shapes without discarding overlapping true-positive bounding boxes.

RQ2: The Critical Impact of Data Curation

The second research question challenged the prevailing „model-centric“ approach in deep learning by questioning the extent to which strategic data curation impacts performance compared to raw data ingestion or aggressive generalized augmentation. The results strongly advocate for a „data-centric“ philosophy in medical and botanical imaging.

Training on the raw, full dataset resulted in severe statistical instability due to extreme class imbalances, causing the models to overfit on under-populated diagnostic fragments (e.g., *Verbascum Endothecia*). Furthermore, early ablation testing with the v13 dataset demonstrated that blindly applying heavy photometric distortions (such as blurring and contrast stretching) aggressively deteriorated model convergence by destroying the delicate microscopic textures the network relies upon. The targeted application of class-consolidation and strictly geometric augmentations in the v17 dataset was the ultimate catalyst for success. This highly curated, pruned dataset directly yielded a 9.9% increase in mAP and a 7.0% increase in mean recall for the YOLOv10x model, conclusively proving

that intelligent data stratification is paramount to unlocking the potential of deep learning in pharmacognosy.

RQ3: Validating Diagnostic Transparency

The final research question addressed the „black-box“ nature of neural networks, asking if Explainable AI (XAI) could verify that the model was utilizing true botanical features rather than overfitting to slide preparation artifacts. The qualitative analysis utilizing spatial detection-density heatmaps and Grad-CAM successfully provided this validation.

The initial dataset analysis revealed a severe operator-induced spatial bias, with a dense concentration of annotations in the exact center of the microscopic slides. However, the spatial heatmaps generated post-training confirmed that the geometric augmentations successfully decoupled the model’s feature extraction from absolute spatial coordinates. More importantly, the Grad-CAM visualizations definitively proved that the YOLOv10x network’s highest confidence scores were structurally anchored to valid biological markers—outlining the specific cellular geometry of plant tissues—rather than focusing on empty slide background, fluid bubbles, or illumination gradients. This diagnostic transparency is essential for the eventual clinical or laboratory adoption of automated pharmacognostic systems.

6.2 Comparison with Prior Work

To properly contextualize the efficacy of the proposed methodological pipeline, the performance of the YOLOv10x architecture is benchmarked against recent, prominent studies utilizing deep learning for microscopic object detection, as previously outlined in Section 2. Table 9 synthesizes these fundamental studies, contrasting their chosen architectures, target domains, and peak mean Average Precision (mAP) metrics against the current study.

Table 9: Comparative analysis of the proposed YOLOv10x model against key prior studies in microscopic object detection.

Study	Architecture	Microscopic Domain	mAP
Pawłowski et al. (2022) [90]	Faster R-CNN	Microbial Species	49.3%
Pang et al. (2025) [97]	DETR (w/ DINO)	Cellular/Nuclear	66.0%
Dahiya et al. (2025) [100]	RF-DETR (w/ DINOv2)	Leukocytes	78.0%
Romphosri et al. (2024) [93]	YOLOv8s	Bacterial Colonies	97.7%
Proposed Study (2026)	YOLOv10x	Pharmaceutical Botany	77.8%

The proposed YOLOv10x model achieved a peak macro mAP@0.50 of 77.8%, situating it highly competitively within the contemporary landscape of microscopic image analysis. It significantly outperforms the foundational Faster R-CNN implementations applied to microbial datasets, such as the 49.3% achieved by Pawłowski et al. Furthermore, the baseline DETR performance observed in this study (67.0%) closely mirrors the 66.0% mAP reported by Pang et al. utilizing a DETR architecture for complex cellular segmentation, validating the general capability ceiling of standard transformer models when deployed on irregular organic structures.

While the 97.7% mAP achieved by Romphosri et al. utilizing the YOLOv8s algorithm appears mathematically superior, this disparity is primarily driven by the inherent morphological differences in the respective datasets. Bacterial colonies (e.g., *E. coli* and *S. aureus*) typically present as highly uniform, well-delineated circular structures against high-contrast

agar backgrounds. Conversely, the pharmaceutical botanical dataset utilized in this study features densely packed, overlapping, and highly irregular geometries (such as fragmented *Althaea* druses and transparent *Verbascum* trichomes). The 77.8% mAP achieved under these severe morphological constraints and bright-field background artifacts highlights the exceptional feature-extraction robustness of the single-stage YOLOv10x paradigm.

6.2.1 Methodological Extensions and Contributions

This research extends existing literature by systematically comparing three distinct architectural paradigms (single-stage, two-stage, and transformer-based) under strictly identical, real-world laboratory conditions. Whereas previous studies, such as Pawłowski et al., relied on synthetically generated data to bypass data scarcity, this study addresses the dataset bottleneck through rigorous manual curation and targeted geometric augmentation.

The substantial 9.9% mAP improvement observed between the raw baseline dataset and the curated v17 dataset quantitatively proves that intelligent data consolidation is as vital as architectural complexity. Ultimately, this work introduces a highly optimized YOLOv10x framework into the novel domain of pharmacognosy, establishing a reliable, NMS-free baseline for future multi-modal botanical analysis that complements and extends current state-of-the-art medical CV applications.

6.3 Practical Implications for Pharmacognosy

The successful empirical validation of the YOLOv10x architecture, augmented by the Explainable AI (XAI) pipeline, carries significant operational implications for the field of pharmaceutical botany. The traditional authentication and quality control of medicinal plants—such as the identification of specific *Verbascum* trichomes or *Althaea officinalis* druses—rely heavily on the manual, subjective expertise of highly trained pharmacognosists. This conventional methodology is inherently vulnerable to human fatigue, inconsistent slide preparation, and low throughput.

The methodological framework established in this thesis provides a foundation for transitioning toward a semi-automated, high-throughput laboratory environment. The practical implications of this shift can be categorized into three primary operational domains:

1. High-Throughput Quality Control and Authentication

The single-stage YOLOv10x architecture is explicitly designed for real-time object detection. Because it bypasses the computationally expensive Non-Maximum Suppression (NMS) step, the inference times are exceptionally low. In a practical laboratory setting, this allows for the rapid, automated screening of digitized slide preparations from incoming batches of raw botanical materials. The system can instantly highlight and tally the presence of crucial diagnostic micro-structures, significantly accelerating the quality control process and ensuring that herbal shipments meet strict pharmacopoeial standards before processing.

2. Standardization and Reduction of Subjective Error

Human microscopic analysis is naturally susceptible to subjective interpretation, particularly when examining highly fragmented or degraded botanical samples. The deployment of a statistically validated neural network introduces a layer of standardized, objective quantification. By utilizing the model as a „second reader“ or diagnostic co-pilot, pharmacognosists can cross-reference their manual findings against the model’s bounding box predictions, thereby minimizing the risk of false negatives (missed target structures) and standardizing the authentication process across different laboratory technicians.

3. Educational Scaffolding via Explainable AI

Beyond direct quality control, the XAI components (Grad-CAM and detection-density heatmaps) possess profound educational utility. Training junior botanists and pharmacy students to identify complex microscopic tissues requires years of guided experience. The Grad-CAM outputs developed in this study effectively serve as automated, localized visual aids. By mathematically highlighting the exact cellular walls and crystalline formations that define a specific class, the system can be utilized as an interactive educational tool, explicitly demonstrating to trainees *where* to focus their attention and *what* morphological characteristics separate valid botanical fragments from laboratory artifacts.

In conclusion, while the neural network is not intended to completely replace the domain expertise of a human pharmacognosist, the proposed YOLOv10x framework serves as a highly reliable, transparent, and rapid computational tool that significantly augments both laboratory throughput and diagnostic consistency.

6.4 Limitations of the Study

While the empirical results robustly validate the use of the YOLOv10x architecture for microscopic botanical analysis, it is imperative to acknowledge the boundaries and inherent limitations of the current experimental framework. These constraints primarily revolve around data scale, domain specificity, and computational demands.

Dataset Scale and Residual Imbalance

Despite the successful implementation of class-consolidation in the v17 dataset, the overall volume of annotated microscopic images remains inherently limited compared to massive generic datasets like MS-COCO or ImageNet. Certain diagnostic micro-structures, such as the *Verbascum Endothecia* or *Cundurango Fiber*, remain statistically under-represented even after pruning. While targeted geometric augmentations mitigated immediate overfitting, the fundamental scarcity of these rare classes restricts the model’s ability to achieve near-perfect recall across the entire botanical spectrum.

Domain and Modality Specificity

The deep learning models were trained exclusively on bright-field microscopy images utilizing specific staining and slide preparation techniques standard to the Department of Pharmacognosy and Pharmaceutical Botany. Consequently, the networks are highly specialized for the morphological profiles of *Verbascum* and *Althaea officinalis* under these exact optical conditions. The models may exhibit degraded performance if subjected to significant domain shifts, such as images acquired via phase-contrast microscopy, fluorescence imaging, or slides exhibiting distinct staining colorations outside the training distribution.

Computational and Architectural Overhead

The training of heavy, state-of-the-art architectures introduced substantial hardware bottlenecks. As documented during the empirical phase, the YOLOv10x and DETR (ResNet-101) models consistently saturated standard GPUs (e.g., NVIDIA T4), requiring migration to high-memory NVIDIA A100 tensors to avoid Out-of-Memory (OOM) failures. Furthermore, the vision transformer paradigm (DETR) demonstrated delayed convergence; its global self-attention mechanisms fundamentally require vastly larger datasets to independently learn spatial priors, rendering it less optimal for this highly specialized, data-constrained application.

6.5 Future Work

The foundational pipeline established in this thesis opens several promising avenues for future research and operational expansion within the intersection of artificial intelligence and pharmaceutical botany.

Generative Data Augmentation

To directly address the limitation of residual class imbalance, future research should explore the application of Generative Artificial Intelligence. Utilizing Generative Adversarial Networks (GANs) or advanced Latent Diffusion Models configured for microscopic imagery could artificially synthesize high-fidelity examples of ultra-rare botanical structures (e.g., *Verbascum Bludgeon*). Injecting these synthetic tensors into the training pipeline could mathematically balance the dataset without requiring years of manual laboratory harvesting and annotation.

Architectural Evolution and Foundation Models

While YOLOv10x proved superior in this study, the rapid evolution of computer vision necessitates continuous architectural evaluation. Future iterations of this research should investigate the application of Vision-Language Foundation Models or specialized medical pre-training paradigms (such as adapting the Segment Anything Model (SAM) for microscopic instances). Furthermore, with the very recent introduction of the fully NMS-free YOLO26 framework, subsequent studies must benchmark the robust YOLOv10x baseline established here against this latest iteration to evaluate potential gains in end-to-end processing speeds and microscopic feature extraction. Additionally, further hyperparameter tuning and significantly extended training schedules for the DETR architecture could potentially unlock its theoretical advantages in processing global morphological contexts, provided the dataset size is concurrently expanded.

Laboratory Integration and Edge Deployment

The ultimate objective of this research trajectory is practical utility. Future development must focus on transitioning the optimized YOLOv10x weights from a Python-based research environment into a user-friendly, graphical laboratory application. Research into model quantization (reducing floating-point precision from FP32 to INT8) or the deployment of lightweight variants (e.g., YOLOv10n) onto Edge AI devices would allow pharmacognosists to run real-time inference directly on standard laboratory desktop computers, bypassing the need for expensive cloud-based GPU clusters.

7 Conclusion

The integration of artificial intelligence into the domain of pharmaceutical botany represents a critical frontier in modern computational biology. This thesis set out to systematically evaluate and optimize state-of-the-art deep learning architectures for the automated detection and classification of highly complex, imbalanced microscopic botanical structures. By directly comparing single-stage (YOLOv10x), two-stage (Faster R-CNN), and transformer-based (DETR) object detection paradigms, this research provides a rigorous empirical framework for transitioning pharmacognostic quality control from subjective manual observation to statistically verifiable, high-throughput automated analysis.

The empirical findings conclusively answered the established research questions. Among the evaluated architectures, the single-stage YOLOv10x model demonstrated statistically significant superiority. By efficiently processing dense, overlapping biological fragments without the bottleneck of Non-Maximum Suppression (NMS), YOLOv10x achieved a peak macro mAP@0.50 of 77.8%, outperforming both Faster R-CNN (67.2%) and DETR (67.0%). The tight 95% Wilson Confidence Intervals further solidified the reliability and generalization capability of this single-stage approach within the specialized microscopic domain.

Crucially, this research highlighted the profound vulnerability of heavy deep learning models to extreme class imbalances and spatial artifacts inherent to raw microscopic slide preparation. The study definitively proved that a „data-centric“ methodology is paramount; strategic dataset pruning, class-consolidation, and targeted geometric augmentation (the v17 dataset) directly catalyzed a 9.9% increase in mean Average Precision and a 7.0% increase in mean recall for the leading architecture. This finding challenges the prevailing reliance on pure algorithmic scaling, demonstrating that intelligent data curation is as vital to model convergence as architectural complexity.

Furthermore, this thesis successfully addressed the persistent „black-box“ limitations of deep neural networks. By engineering a custom Explainable AI (XAI) pipeline utilizing spatial detection-density heatmaps and surrogate-driven Grad-CAM visualizations, the research provided essential diagnostic transparency. The qualitative analysis verified that the YOLOv10x network’s highest confidence predictions were legitimately anchored to true morphological characteristics—such as the complex cellular walls of *Althaea officinalis* druses and the crystalline geometries of *Verbascum* trichomes—rather than overfitting to operator-induced centering biases or bright-field illumination anomalies.

In summary, this thesis establishes a highly optimized, transparent, and reproducible computational pipeline for microscopic botanical authentication. While challenges regarding the scarcity of ultra-rare diagnostic classes remain, the foundational success of the YOLOv10x framework paves the way for the deployment of real-time, AI-assisted diagnostic co-pilots within the Department of Pharmacognosy and Pharmaceutical Botany. By augmenting human expertise with automated, scalable machine vision, this research contributes a vital technological stepping stone toward the future of standardized, high-throughput pharmaceutical quality control.

References

- [1] Vincent, J. DeepMind’s Demis Hassabis on the future of AI. *The Verge*, July 2021, online interview with Demis Hassabis. Available from:
<https://www.theverge.com/2021/7/22/22587593/deepmind-demis-hassabis-alphafold-2-ai-future-interview>
- [2] Archit, A.; Freckmann, L.; Nair, S.; et al. Segment Anything for Microscopy. *Nature Methods*, volume 22, Feb. 2025: pp. 579–591, doi:10.1038/s41592-024-02580-4.
- [3] Lipp, P.; Fornasiero, E. F.; Valitutti, S.; et al. Highly adaptable deep-learning platform for automated detection and analysis of vesicle exocytosis. *Nature Communications*, volume 16, 2025, doi:10.1038/s41467-025-61579-3. Available from:
<https://www.nature.com/articles/s41467-025-61579-3>
- [4] Zhang, Y.; Lu, H.; Li, Y.; et al. The Development of Microscopic Imaging Technology and its Application in Micro- and Nanotechnology. *Frontiers in Chemistry*, volume 10, July 2022: p. 931169, ISSN 2296-2646, doi:10.3389/fchem.2022.931169. Available from:
<https://www.frontiersin.org/articles/10.3389/fchem.2022.931169>
- [5] Wang, Y.; Liao, R.; Dai, J.; et al. Differentiation of suspended particles by polarized light scattering at 120°. *Optics Express*, volume 26, no. 17, Aug. 2018: pp. 22419–22431, ISSN 1094-4087, doi:10.1364/OE.26.022419. Available from:
<https://opg.optica.org/oe/abstract.cfm?uri=oe-26-17-22419>
- [6] Wollman, A. J. M.; Nudd, R.; Hedlund, E. G.; et al. From Animaculum to Single Molecules: 300 Years of the Light Microscope. *Open Biology*, volume 5, no. 4, 2015: p. 150019, doi:10.1098/rsob.150019, a comprehensive historical review tracing the evolution of the light microscope from early 17th-century observations of “animalcules” by pioneers such as van Leeuwenhoek and Hooke, through major technological milestones, to modern single-molecule imaging techniques in the 21st century. Highlights the interplay between optical innovation, scientific discovery, and imaging resolution over three centuries.
- [7] Noble, C. A.; Biesemier, A. P.; McClees, S. F.; et al. The history of the microscope reflects advances in science and medicine. *Seminars in Diagnostic Pathology*, volume 41, no. 3, 2024: pp. 206–212, ISSN 0740-2570, doi:10.1053/j.semmp.2024.01.002, review article.
- [8] Giepmans, B. N. G.; Taatjes, D. J.; Wolstencroft, K. J. In focus: data management and data analysis in microscopy. *Histochemistry and Cell Biology*, volume 160, no. 2, aug 2023: pp. 165–167, doi:10.1007/s00418-023-02226-0. Available from:
<https://doi.org/10.1007/s00418-023-02226-0>
- [9] Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, volume 3, mar 2016: p. 160018, doi:10.1038/sdata.2016.18. Available from:
<https://doi.org/10.1038/sdata.2016.18>

- [10] Usman, S.; Mehmood, R.; Katib, I.; et al. Data Locality in High Performance Computing, Big Data, and Converged Systems: An Analysis of the Cutting Edge and a Future System Architecture. *Applied Sciences*, volume 12, no. 1, Dec. 2022: p. 53, ISSN 2079-9292, doi:10.3390/app12010053. Available from: <https://www.mdpi.com/2079-9292/12/1/53>
- [11] Modha, D. S.; Akopyan, F.; Andreopoulos, A.; et al. Neural inference at the frontier of energy, space, and time. *Science*, volume 382, no. 6668, Oct. 2023: pp. 329–335, doi:10.1126/science.adh1174.
- [12] Hennessy, J. L.; Patterson, D. A. A new golden age for computer architecture. *Communications of the ACM*, volume 62, no. 2, jan 2019: pp. 48–60, doi:10.1145/3282260. Available from: <https://doi.org/10.1145/3282260>
- [13] Chien, A. A. Computer Architecture: Disruption from Above. *Communications of the ACM*, volume 61, no. 9, Sept. 2018: p. 5, doi:10.1145/3243136.
- [14] Yazdanbakhsh, A. Beyond Moore’s Law: Harnessing the Redshift of Generative AI with Effective Hardware–Software Co–Design. *arXiv preprint arXiv:2504.06531*, Apr. 2025.
- [15] Clark, S. The End of Moore’s Law? AI Chipmakers Say It’s Already Happened. Mar. 2025, accessed: 2025-08-06. Available from: <https://www.vktr.com/ai-technology/the-end-of-moores-law-ai-chipmakers-say-its-already-happened/>
- [16] Ihde, N.; Marten, P.; Eleliemy, A.; et al. A Survey of Big Data, High Performance Computing, and Machine Learning Benchmarks. Technical report, Hasso Plattner Institute, University of Potsdam, 2021. Available from: https://hpi.de/fileadmin/user_upload/fachgebiete/rabl/publications/2021/A_Survey_of_Big_Data_High_Performance_Computing_and_Machine_Learning_Benchmarks.pdf
- [17] Lu, M. Y.; Chen, B.; Williamson, D. F. K.; et al. A Multimodal Generative AI Copilot for Human Pathology. *Nature*, volume 634, June 2024: pp. 466–473, doi:10.1038/s41586-024-07618-3.
- [18] Kumar, Y.; Marchena, J.; Awlla, A. H.; et al. The AI-Powered Evolution of Big Data. *Applied Sciences*, volume 14, no. 22, nov 2024: p. 10176, ISSN 2076-3417, doi:10.3390/app142210176. Available from: <https://www.mdpi.com/2076-3417/14/22/10176>
- [19] Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, volume 60, no. 6, May 2017: pp. 84–90, doi:10.1145/3065386. Available from: <https://doi.org/10.1145/3065386>
- [20] Al-Iqubaydhi, N.; Alenezi, A.; Alanazi, T.; et al. Deep Learning for Unmanned Aerial Vehicles Detection: A Review. *Journal of Systems Architecture*, volume 145, 2024: p. 102828, doi:10.1016/j.sysarc.2023.102828. Available from: <https://www.sciencedirect.com/science/article/abs/pii/S1574013723000813>

- [21] Qiao, Y.; Yin, J.; Wang, W.; et al. Survey of Deep Learning for Autonomous Surface Vehicles in Marine Environments. *IEEE Transactions on Intelligent Transportation Systems*, volume 24, no. 4, Apr. 2023: pp. 3678–3701, doi:10.1109/TITS.2023.3235911.
- [22] Zhang, C.; Sjarif, N. N. A.; Ibrahim, R. Deep learning models for price forecasting of financial time series: A review of recent advancements: 2020–2022. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Sept. 2023, doi:10.1002/widm.1519, advanced Review; Edited by Witold Pedrycz. Available from: <https://doi.org/10.1002/widm.1519>
- [23] Mienye, E.; Jere, N.; Obaido, G.; et al. Deep Learning in Finance: A Survey of Applications and Techniques, Aug. 2024, doi:10.20944/preprints202408.1365.v1, preprint, not peer-reviewed. Available from: <https://www.preprints.org/manuscript/202408.1365/v1>
- [24] Sahu, S. K.; Mokhade, A.; Bokde, N. D. An Overview of Machine Learning, Deep Learning, and Reinforcement Learning-Based Techniques in Quantitative Finance: Recent Progress and Challenges. *Applied Sciences*, volume 13, no. 3, Feb. 2023: p. 1956, doi:10.3390/app13031956. Available from: <https://doi.org/10.3390/app13031956>
- [25] Wang, A.; Chen, H.; Liu, L.; et al. YOLOv10: Real-Time End-to-End Object Detection. 2024, doi:arXiv:2405.14458, arXiv preprint arXiv:2405.14458. Available from: <https://arxiv.org/abs/2405.14458>
- [26] Zhao, Y.; Lv, W.; Xu, S.; et al. DETRs Beat YOLOs on Real-time Object Detection. arXiv preprint arXiv:2304.08069, 2023. Available from: <https://arxiv.org/abs/2304.08069>
- [27] Magdy, A.; Moustafa, M. S.; Ebied, H. M.; et al. Lightweight Faster R-CNN for Object Detection in Optical Remote Sensing Images. *Scientific Reports*, volume 15, 2025, doi:10.1038/s41598-025-99242-y.
- [28] Hung, J.; Stefanie C. P. L.; Nery, O. A.; et al. Applying Faster R-CNN for Object Detection on Malaria Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 56–61.
- [29] Zhang, C.; Yue, J.; Fu, J.; et al. River floating object detection with transformer model in real time. *Scientific Reports*, volume 15, 2025: p. 9026, doi:10.1038/s41598-025-9026-1, open access.
- [30] Zhou, D.; Xu, H.; Liu, W.; et al. LN-DETR: cross-scale feature fusion and re-weighting for lung nodule detection. *Scientific Reports*, volume 15, 2025: p. 15543, doi:10.1038/s41598-025-15543-4, open access.
- [31] Zhang, X.; Sun, W.; Chen, K.; et al. A novel method for intelligent operation and maintenance of transformers using deep visual large model DETR+X and digital twin. *Scientific Reports*, volume 15, 2025: p. 98, doi:10.1038/s41598-024-80553-1, open access.

- [32] Nature Methods Editorial Board. Calling all data. *Nature Methods*, volume 22, 2025: p. 1387, doi:10.1038/s41592-025-0233-3, editorial.
- [33] Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast). Official Journal of the European Union, L 172/56, 26 June 2019, June 2019, doi:10.3000/19770677.L_2019.172.eng, open Data Directive (recast); setting minimum rules for EU Member States on public sector information reuse. Available from: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32019L1024>
- [34] Swedlow, J. R.; Kankaanpää, P.; Sarkans, U.; et al. A global view of standards for open image data formats and repositories. *Nature Methods*, volume 18, no. 12, 2021: pp. 1440–1446, doi:10.1038/s41592-021-01113-7.
- [35] European Bioinformatics Institute (EMBL-EBI). BioImage Archive. <https://www.ebi.ac.uk/bioimage-archive/>, 2022, accessed: 2025-08-14; Open-access repository for biological and microscopy image datasets.
- [36] Giner-Miguelez, J.; Gómez, A.; Cabot, J. On the Readiness of Scientific Data Papers for a Fair and Transparent Use in Machine Learning. *Scientific data*, volume 12, no. 1, 2025: p. 61.
- [37] Lu, S.-Y.; Wang, S.-H.; Zhang, Y.-D. SAFNet: A deep spatial attention network with classifier fusion for breast cancer detection. *Computers in Biology and Medicine*, volume 148, Sept. 2022: p. 105812, doi:10.1016/j.compbiomed.2022.105812, epub 2022 Jul 8. PMID: 35834967. Available from: <https://doi.org/10.1016/j.compbiomed.2022.105812>
- [38] Zhou, J.; Luo, L.-Y.; Dou, Q.; et al. Weakly supervised 3D deep learning for breast cancer classification and localization of the lesions in MR images. *Journal of Magnetic Resonance Imaging*, volume 50, no. 4, Oct. 2019: pp. 1144–1151, doi:10.1002/jmri.26721, epub 2019-03-29. The first two authors contributed equally to this work.
- [39] Chai, B.; Efstathiou, C.; Yue, H.; et al. Opportunities and challenges for deep learning in cell dynamics research. *Trends in Cell Biology*, volume 34, no. 11, 2024: pp. 801–814, doi:10.1016/j.tcb.2023.10.010, accessed: 2025-08-14. Available from: <https://www.sciencedirect.com/science/article/pii/S0962892423002283>
- [40] ud din, N. M.; Dar, R. A.; Rasool, M.; et al. Breast cancer detection using deep learning: Datasets, methods, and challenges ahead. *Computers in Biology and Medicine*, volume 149, 2022: p. 106073, ISSN 0010-4825, doi:https://doi.org/10.1016/j.compbiomed.2022.106073. Available from: <https://www.sciencedirect.com/science/article/pii/S0010482522007818>
- [41] Spiers, H.; Songhurst, H.; Nightingale, L.; et al. Deep learning for automatic segmentation of the nuclear envelope in electron microscopy data, trained with volunteer segmentations. *Traffic*, volume 22, no. 7, 2021: pp. 240–253.
- [42] Munappy, A. R.; Bosch, J.; Olsson, H. H.; et al. Data management for production quality deep learning models: Challenges and solutions. *Journal of Systems and Software*, volume 191, 2022: p. 111359.

- [43] Rajpurkar, P.; Chen, E.; Banerjee, O.; et al. AI in health and medicine. *Nature medicine*, volume 28, no. 1, 2022: pp. 31–38.
- [44] Li, X.; Zhang, Y.; Wu, J.; et al. Challenges and opportunities in bioimage analysis. *Nature Methods*, volume 20, no. 7, Jul 2023: pp. 958–961, doi:10.1038/s41592-023-01900-4.
- [45] Rani, P.; Kotwal, S.; Manhas, J.; et al. Machine learning and deep learning based computational approaches in automatic microorganisms image recognition: methodologies, challenges, and developments. *Archives of Computational Methods in Engineering*, volume 29, no. 3, 2022: pp. 1801–1837.
- [46] Shi, L. Enhancing medical explainability in deep learning for age-related macular degeneration diagnosis. *Scientific Reports*, volume 15, 2025: p. 16975, doi:10.1038/s41598-025-01496-z. Available from: <https://doi.org/10.1038/s41598-025-01496-z>
- [47] De Vries, B. M.; Zwezerijnen, G. J.; Burchell, G. L.; et al. Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review. *Frontiers in medicine*, volume 10, 2023: p. 1180773.
- [48] Adam, M.; AlEisa, H.; Zanin, S.; et al. Advanced object detection for smart accessibility: a Yolov10 with marine predator algorithm to aid visually challenged people. *Scientific Reports*, volume 15, 2025: p. 20759, doi:10.1038/s41598-025-04959-5, published: 01 July 2025; Received: 13 February 2025; Accepted: 29 May 2025. Available from: <https://www.nature.com/articles/s41598-025-04959-5>
- [49] Moreni, M.; Theau, J.; Foucher, S. Do you get what you see? Insights of using mAP to select architectures of pretrained neural networks for automated aerial animal detection. *PLoS One*, volume 18, no. 4, 2023: p. e0284449, doi:10.1371/journal.pone.0284449.
- [50] Wen, J.; Liu, H.; Li, J. A Task-Risk Consistency Object Detection Framework Based on Deep Reinforcement Learning. *Remote Sensing*, volume 15, no. 20, 2023: p. 5031, doi:10.3390/rs15205031.
- [51] Chun, Y.; Hua, X.; Qi, C.; et al. Improved marine predators algorithm for engineering design optimization problems. *Scientific Reports*, 2024, doi:10.1038/s41598-024-63826-x, includes detailed formulation of Elite and Prey matrices and optimization phases (HVR, UVR, LVR).
- [52] Houssein, E. H.; Hassaballah, M.; Ibrahim, I. E.; et al. An automatic arrhythmia classification model based on improved Marine Predators Algorithm and Convolutions Neural Networks. *Expert Systems with Applications*, volume 187, 2022: p. 115936, ISSN 0957-4174, doi:<https://doi.org/10.1016/j.eswa.2021.115936>. Available from: <https://www.sciencedirect.com/science/article/pii/S0957417421012902>
- [53] Liu, F.; Dong, Z.; Gorbani, B. Utilizing deep belief network optimized by balanced Manta ray foraging optimization algorithm for estimating the shear Wall's shear strength. *Scientific Reports*, volume 15, 2025: p. 7443,

- doi:10.1038/s41598-025-91466-2. Available from:
<https://www.nature.com/articles/s41598-025-91466-2>
- [54] Qin, T.; Yan, G.; Jiang, H.; et al. Optimizing deep belief network for concrete crack detection via a modified design of ideal gas molecular dynamics. *Scientific Reports*, volume 15, 2025: p. 9070, doi:10.1038/s41598-025-93397-4. Available from:
<https://www.nature.com/articles/s41598-025-93397-4>
 - [55] Volosevici, D.; Isbasoiu, G. D. Surveillance as a Socio-Technical System: Behavioral Impacts and Self-Regulation in Monitored Environments. *Systems*, volume 13, no. 7, 2025: p. 614, doi:10.3390/systems13070614, this article belongs to the Section Systems Practice in Social Science. Available from:
<https://doi.org/10.3390/systems13070614>
 - [56] Lin, L.; Purnell, N. A World With a Billion Cameras Watching You Is Just Around the Corner. *The Wall Street Journal*, Dec. 2019, world Section. Available from:
<https://www.wsj.com/articles/a-billion-surveillance-cameras-forecast-to-be-watching-within-two-years-11575565402>
 - [57] Peterson, J. K.; Densley, J. A.; Pyrooz, D. C. Mental health consequences of exposure to mass and non-mass shootings in a national sample of US adults. *Nature Mental Health*, volume 3, Apr. 2025: pp. 530–537, open access.
 - [58] Khan, S.; Fuadah, S.; Ikram, N.; et al. Exploring mental health challenges and violence risks in high school population. *Scientific Reports*, volume 15, July 2025: p. 22367, open access; Article number: 22367.
 - [59] Xu, G.; Li, T.; Li, L. Food insecurity is associated with violent behaviors among in-school adolescents. *Scientific Reports*, volume 15, May 2025: p. 17526, open access; Article number: 17526.
 - [60] Ciampiconi, L.; Elwood, A.; Leonardi, M.; et al. A survey and taxonomy of loss functions in machine learning. *arXiv preprint arXiv:2301.05579*, 2023, doi:10.48550/arXiv.2301.05579, version 2, last revised 18 Nov 2024. Available from:
<https://doi.org/10.48550/arXiv.2301.05579>
 - [61] Abhishek, A.; Jha, R. K.; Sinha, R.; et al. Automated detection and classification of leukemia on a subject-independent test dataset using deep transfer learning supported by Grad-CAM visualization. *Biomedical Signal Processing and Control*, volume 83, 2023: p. 104722, ISSN 1746-8094, doi:10.1016/j.bspc.2023.104722.
 - [62] Wang, X.; Yang, W.; Qi, W.; et al. STarNet: A spatio-temporal and Riemannian network for high-performance motor imagery decoding. *Neural Networks*, volume 178, 2024: p. 106471, ISSN 0893-6080, doi:https://doi.org/10.1016/j.neunet.2024.106471. Available from:
<https://www.sciencedirect.com/science/article/pii/S0893608024003952>
 - [63] Hajizadeh, M.; Sabokrou, M.; Rahmani, A. STARNet: spatio-temporal aware recurrent network for efficient video object detection on embedded devices. *Machine Vision and Applications*, volume 35, no. 2, 2024: p. 23.

- [64] Zakariah, M.; Alnuaim, A. Recognizing human activities with the use of Convolutional Block Attention Module. *Egyptian Informatics Journal*, volume 27, 2024: p. 100536, ISSN 1110-8665, doi:<https://doi.org/10.1016/j.eij.2024.100536>. Available from: <https://www.sciencedirect.com/science/article/pii/S1110866524000999>
- [65] Woo, S.; Park, J.; Lee, J.-Y.; et al. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
- [66] Zhao, Q.; Liu, S.; Zhang, S.; et al. Campus risk detection using the S-YOLOv10-SIC network and a self-calibrated illumination algorithm. *Scientific Reports*, volume 15, 2025: p. 24209, doi:10.1038/s41598-025-24209-4. Available from: <https://doi.org/10.1038/s41598-025-24209-4>
- [67] Zhang, Y.; Jia, N. A target detection model HR-YOLO for advanced driver assistance systems in foggy conditions. *Scientific Reports*, volume 15, no. 1, 2025: p. 13067.
- [68] Wang, Y.; Wang, F.; Qin, Z.; et al. A non-invasive diagnostic approach for neuroblastoma utilizing preoperative enhanced computed tomography and deep learning techniques. *Scientific Reports*, volume 15, no. 1, 2025: p. 14652.
- [69] Qiao, L.; Veltrup, M.; Mayer, B. Comparison of mask R-CNN and YOLOv8-seg for improved monitoring of the PCB surface during laser cleaning. *Scientific Reports*, volume 15, no. 1, 2025: p. 17185.
- [70] Hong, J.; Ye, K.; Qiu, S. Study on lightweight strategies for L-YOLO algorithm in road object detection. *Scientific Reports*, volume 15, no. 1, 2025: p. 7649.
- [71] Leng, Y.; Liu, J. Improved faster R-CNN for steel surface defect detection in industrial quality control. *Scientific Reports*, volume 15, no. 1, 2025: p. 30093.
- [72] Ferrante, G. S.; Vasconcelos Nakamura, L. H.; Sampaio, S.; et al. Evaluating YOLO architectures for detecting road killed endangered Brazilian animals. *Scientific reports*, volume 14, no. 1, 2024: p. 1353.
- [73] Gui, H.; Jiao, H.; Li, L.; et al. Breast tumor detection and diagnosis using an improved faster R-CNN in dce-MRI. *Bioengineering*, volume 11, no. 12, 2024: p. 1217.
- [74] Li, H.; Gu, Y.; Han, J.; et al. Faster R-CNN-MobileNetV3 Based Micro Expression Detection for Autism Spectrum Disorder. *AI Medicine*, 2025: pp. 2–2.
- [75] Ying, H.; Liu, X.; Zhang, M.; et al. A multicenter clinical AI system study for detection and diagnosis of focal liver lesions. *Nature Communications*, volume 15, no. 1, 2024: p. 1131.
- [76] Dong, X.; Bao, J.; Chen, D.; et al. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12124–12134.
- [77] Li, F.; Yan, H.; Shi, L. Multi-scale coupled attention for visual object detection. *Scientific Reports*, volume 14, no. 1, 2024: p. 11191.

- [78] Khan, S.; Naseer, M.; Hayat, M.; et al. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, volume 54, no. 10s, 2022: pp. 1–41.
- [79] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [80] Carion, N.; Massa, F.; Synnaeve, G.; et al. End-to-end object detection with transformers. In *European conference on computer vision*, Springer, 2020, pp. 213–229.
- [81] He, W.; Zhang, Y.; Xu, T.; et al. Object detection for medical image analysis: Insights from the RT-DETR model. In *Proceedings of the 2025 International Conference on Artificial Intelligence and Computational Intelligence*, 2025, pp. 415–420.
- [82] Liu, Z.; Lin, Y.; Cao, Y.; et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.
- [83] Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [84] Xu, Y.; Shen, Y.; Fernandez-Granda, C.; et al. Understanding differences in applying detr to natural and medical images. *arXiv preprint arXiv:2405.17677*, 2024.
- [85] Hu, Y.; Chen, N.; Hou, Y.; et al. Lightweight deep learning for real-time road distress detection on mobile devices. *Nature Communications*, volume 16, no. 1, 2025: p. 4212.
- [86] Shi, Y.; Jia, Y.; Zhang, X. FocusDet: an efficient object detector for small object. *Scientific Reports*, volume 14, no. 1, 2024: p. 10697.
- [87] Ghahremani, M.; Ernhof, B. R.; Wang, J.; et al. Organ-DETR: Organ Detection via Transformers. *IEEE Transactions on Medical Imaging*, 2025.
- [88] Archi, A.; Freckmann, L.; Nair, S.; et al. Segment anything for microscopy. *Nature Methods*, volume 22, no. 3, 2025: pp. 579–591.
- [89] Hardo, G.; Li, R.; Bakshi, S. Quantitative microbiology with widefield microscopy: navigating optical artefacts for accurate interpretations. *npj Imaging*, volume 2, no. 1, 2024: p. 26.
- [90] Pawłowski, J.; Majchrowska, S.; Golan, T. Generation of microbial colonies dataset with deep learning style transfer. *Scientific Reports*, volume 12, no. 1, 2022: p. 5212.
- [91] Zhao, Y.; Sun, F.; Wu, X. FEB-YOLOv8: A multi-scale lightweight detection model for underwater object detection. *PLoS One*, volume 19, no. 9, 2024: p. e0311173.
- [92] Liu, S.; Qi, L.; Qin, H.; et al. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759–8768.

- [93] Romphosri, S.; Pissuwan, D.; Wattanavichean, N.; et al. Rapid alignment-free bacteria identification via optical scattering with LEDs and YOLOv8. *Scientific Reports*, volume 14, no. 1, 2024: p. 20498.
- [94] Zhang, H.; Li, F.; Liu, S.; et al. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.
- [95] Greenwald, N. F.; Miller, G.; Moen, E.; et al. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nature biotechnology*, volume 40, no. 4, 2022: pp. 555–565.
- [96] Pachitariu, M.; Stringer, C. Cellpose 2.0: how to train your own model. *Nature methods*, volume 19, no. 12, 2022: pp. 1634–1641.
- [97] Pang, M.; Roy, T. K.; Wu, X.; et al. CelloType: a unified model for segmentation and classification of tissue images. *Nature methods*, volume 22, no. 2, 2025: pp. 348–357.
- [98] Połap, D.; Woźniak, M. Red fox optimization algorithm. *Expert Systems with Applications*, volume 166, 2021: p. 114107.
- [99] Dwyer, B. and Nelson, J. and Hansen, T. et al. Roboflow.
url<https://roboflow.com>, 2024, accessed: September 12, 2025.
- [100] Dahiya, N.; Prakash, D.; Kundu, S.; et al. Optimised RFO tuned RF-DETR model for precision urine microscopy for renal and systemic disease diagnosis. *Scientific Reports*, volume 15, no. 1, 2025: p. 25842.
- [101] Kalluri, P. R.; Agnew, W.; Cheng, M.; et al. Computer-vision research powers surveillance technology. *Nature*, 2025: pp. 1–7.
- [102] Schulze Buschoff, L. M.; Akata, E.; Bethge, M.; et al. Visual cognition in multimodal large language models. *Nature Machine Intelligence*, volume 7, no. 1, 2025: pp. 96–106.
- [103] Sun, Z.; Chen, Y.-J.; Yang, Y.-H.; et al. Machine learning modelling for multi-order human visual motion processing. *Nature Machine Intelligence*, 2025: pp. 1–16.
- [104] Alpaydin, E. *Introduction to Machine Learning*. Adaptive Computation and Machine Learning, Cambridge, MA; London, England: The MIT Press, third edition, 2014, ISBN 978-0-262-02818-9.
- [105] Kelleher, J. D. *Deep Learning*. The MIT Press Essential Knowledge Series, Cambridge, Massachusetts; London, England: The MIT Press, 2019, ISBN 9780262537551, paperback (pbk. : alk. paper).
- [106] Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*. Cambridge, MA: The MIT Press, 2016, book and supplementary material (see website). Available from: <https://www.deeplearningbook.org>
- [107] Hutter, F.; Kotthoff, L.; Vanschoren, J. (editors). *Automated Machine Learning: Methods, Systems, Challenges*. The Springer Series on Challenges in Machine Learning, Cham, Switzerland: Springer Nature Switzerland AG, 2019, ISBN

- 978-3-030-05317-8, doi:10.1007/978-3-030-05318-5, open access; corrected publication 2019. Available from: <https://doi.org/10.1007/978-3-030-05318-5>
- [108] Oracle Corporation. Physical Database Limits. <https://docs.oracle.com/en/database/oracle/oracle-database/19/refrn/physical-database-limits.html>, 2019, part of Oracle Database 19c Reference. Accessed 19 September 2025. Available from: <https://docs.oracle.com/en/database/oracle/oracle-database/19/refrn/physical-database-limits.html>
 - [109] Jung, A. *Machine Learning: The Basics*. Machine Learning: Foundations, Methodologies, and Applications, Singapore: Springer Nature Singapore Pte Ltd., 2022, ISBN 978-981-16-8192-9, doi:10.1007/978-981-16-8193-6, series editors: Kay Chen Tan and Dacheng Tao.
 - [110] Balaram, V.; Sawant, S. Indicator minerals, pathfinder elements, and portable analytical instruments in mineral exploration studies. *Minerals*, volume 12, no. 4, 2022: p. 394.
 - [111] Buduma, N.; Buduma, N.; Papa, J. *Fundamentals of Deep Learning*. O'Reilly Media, 2022, ISBN 9781492082156. Available from: <https://books.google.cz/books?id=2-5vEAAAQBAJ>
 - [112] Azevedo, F. A.; Carvalho, L. R.; Grinberg, L. T.; et al. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *Journal of Comparative Neurology*, volume 513, no. 5, 2009: pp. 532–541.
 - [113] Smith, K. What’s so special about the human brain? *Nature*, October 2024, doi:10.1038/d41586-024-03425-y, immersive article. Available from: <https://www.nature.com/immersive/d41586-024-03425-y/index.html>
 - [114] Nielsen, M. A. *Neural Networks and Deep Learning*. Determination Press, 2015, <http://neuralnetworksanddeeplearning.com/>.
 - [115] Venkatesan, R.; Li, B. *Convolutional Neural Networks in Visual Computing: A Concise Guide*. Data-Enabled Engineering, Boca Raton, FL: CRC Press, 2018, ISBN 978-1-4987-7039-2, series Editor: Nong Ye.
 - [116] Zhang, A.; Lipton, Z. C.; Li, M.; et al. *Dive into Deep Learning*. Cambridge University Press, 2023, <https://d2l.ai>.
 - [117] Girshick, R.; Donahue, J.; Darrell, T.; et al. Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 38, no. 1, 2016: pp. 142–158, doi:10.1109/TPAMI.2015.2437384.
 - [118] Szeliski, R. *Computer Vision: Algorithms and Applications*. Cham: Springer, second edition, 2022, ISBN 978-3030343712, doi:10.1007/978-3-030-34372-9. Available from: <https://szeliski.org/Book/>
 - [119] Redmon, J.; Divvala, S.; Girshick, R.; et al. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

- [120] Terven, J. R.; Córdova-Esparza, D.-M. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Machine Learning and Knowledge Extraction*, volume 5, no. 4, 2023: pp. 1680–1716.
- [121] Wang, C.-Y.; Liao, H.-Y. M.; Wu, Y.-H.; et al. CSPNet: A New Backbone that can Enhance Learning Capability of CNN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020, pp. 390–391.
- [122] Liu, S.; Qi, L.; Qin, H.; et al. Path Aggregation Network for Instance Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8759–8768.
- [123] Yeerjiang, A.; Wang, Z.; Huang, X.; et al. YOLOv1 to YOLOv10: A Comprehensive Review of YOLO Variants and Their Application in Medical Image Detection. *Journal of Artificial Intelligence Practice*, volume 7, 2024: pp. 112–122.
- [124] Chakrabarty, S. YOLO26: An Analysis of NMS-Free End to End Framework for Real-Time Object Detection. 2026, 2601.12882. Available from: <https://arxiv.org/abs/2601.12882>
- [125] Shehzadi, T.; et al. Object Detection with Transformers: A Review. *Sensors*, volume 25, no. 19, 2025: p. 6025.
- [126] Zhu, X.; Su, W.; Lu, L.; et al. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *International Conference on Learning Representations*, 2021.
- [127] Lv, W.; Zhao, Y.; Xu, S.; et al. DETRs Beat YOLOs on Real-time Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1–10.
- [128] Lin, T.-Y.; Maire, M.; Belongie, S.; et al. Microsoft COCO: Common objects in context. In *European conference on computer vision*, Springer, 2014, pp. 740–755.
- [129] Everingham, M.; Van Gool, L.; Williams, C. K.; et al. The pascal visual object classes (VOC) challenge. *International journal of computer vision*, volume 88, no. 2, 2010: pp. 303–338.
- [130] Padilla, R.; Netto, S. L.; Da Silva, E. A. A survey on performance metrics for object-detection algorithms. In *2020 international conference on systems, signals and image processing (IWSSIP)*, IEEE, 2020, pp. 237–242.
- [131] Wilson, E. B. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, volume 22, no. 158, 1927: pp. 209–212.
- [132] Brown, L. D.; Cai, T. T.; DasGupta, A. Interval estimation for a binomial proportion. *Statistical science*, volume 16, no. 2, 2001: pp. 101–117.
- [133] Agresti, A.; Coull, B. A. Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician*, volume 52, no. 2, 1998: pp. 119–126.

- [134] Agresti, A. *Categorical data analysis*. John Wiley & Sons, third edition, 2013.
- [135] Hassan, M.; et al. Towards Transparent AI in Medical Imaging: Fracture Detection in Hand Radiographs with Grad-CAM Insights. *SciTePress*, 2025.
- [136] Al-Antari, M. A.; et al. Analyzing explainability of YOLO-based breast cancer detection using heat map visualizations. *Quantitative Imaging in Medicine and Surgery*, volume 14, no. 8, 2024: p. 139112.
- [137] Selvaraju, R. R.; Cogswell, M.; Das, A.; et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
- [138] Kirchknopf, A.; Slijepčević, D.; Zeppelzauer, M. Explaining YOLO: Leveraging Grad-CAM to Explain Object Detections. *arXiv preprint arXiv:2211.12130*, 2022.
- [139] Roboflow. Roboflow — computer vision platform for dataset annotation, preprocessing and export. <https://roboflow.com>, 2025, accessed: 2025-10-14.
- [140] Google LLC. Google Colab — hosted Jupyter notebook environment for interactive computing and GPU/TPU access. <https://colab.research.google.com>, 2025, accessed: 2025-10-14.
- [141] NVIDIA Corporation. NVIDIA Tesla T4 Tensor Core GPU — Product Brief. <https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/tesla-t4/t4-tensor-core-product-brief.pdf>, 2018, accessed: 2025-10-14.
- [142] NVIDIA Corporation. NVIDIA A100 Tensor Core GPU — Data Sheet. <https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-us-nvidia-1758950-r4-web.pdf>, 2021, accessed: 2025-10-14.
- [143] Li, X.; et al. TCMP-300: A Comprehensive Traditional Chinese Medicinal Plant Dataset for Plant Recognition. *Scientific Data*, volume 12, no. 1, 2025: p. 243.
- [144] Asenov, A.; et al. Pharmacognostic analysis and teachable machine automated identification of *Uvae ursi folium*, *Vitis idaeae folium* and *Buxi folium*. *Journal of Pharmacy & Pharmacognosy Research*, volume 13, no. 4, 2025: pp. 1191–1204.
- [145] Bisong, E.; Bisong, E. *Google Colaboratory*. Springer, 2019, 59–64 pp.
- [146] NVIDIA Corporation. NVIDIA A100 Tensor Core GPU Architecture. 2020, whitepaper. Accessed: 2026-03-09. Available from: <https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf>
- [147] Paszke, A.; Gross, S.; Massa, F.; et al. PyTorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, volume 32, 2019.
- [148] Zheng, Z.; Wang, P.; Liu, W.; et al. Distance-IoU loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 2020, pp. 12993–13000.

- [149] Zhuang, F.; Qi, Z.; Duan, K.; et al. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, volume 109, no. 1, 2020: pp. 43–76.
- [150] Ullah, Z.; et al. Transfer learning for medical image analysis: A comprehensive review. *Artificial Intelligence in Medicine*, volume 148, 2024: p. 102763.
- [151] Wu, Y.; Kirillov, A.; Massa, F.; et al. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019.
- [152] Ultralytics. Ultralytics YOLO Documentation. <https://docs.ultralytics.com/>, 2024.
- [153] Wolf, T.; et al. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.
- [154] Falcon, W.; et al. PyTorch Lightning. *GitHub. Note:* <https://github.com/PyTorchLightning/pytorch-lightning>, volume 3, no. 6, 2019.
- [155] Mumtaz, A.; et al. Medical Image Data Augmentation Techniques: A Comprehensive Review. *IEEE Access*, volume 12, 2024: pp. 14567–14589.

Glossary

AI artificial intelligence

AODSA-MPAVCP advanced object detection for smart accessibility using the marine predator algorithm to aid visually challenged people

BD big data

CBAM convolutional block attention module

CNNs convolutional neural networks

ConvNet convolutional neural network

CSP cross stage partial

CSWin cross-shaped window

CT computed tomography

CV computer vision

DBN deep belief network

DETR DETection TRansformer

DINO DETR with improved de-noising anchor boxes

DL deep learning

DNN deep neural network

DQM dense query matching

DTs decision trees

EU European Union

FAIR findable accessible interoperable and reusable

FLOPS floating point operations per second

FNN feed-forward neural network

FP false positive

FPN feature pyramid network

GPUs graphics processing units

Grad-CAM gradient-weighted class activation mapping

HPC high performance computing

HPO hyperparameter optimization

HW hardware

IoU intersection over union

LePE locally-enhanced positional encoding

LiAIDS liver artificial intelligence diagnosis system

mAP mean average precision

mAR mean average recall

ML machine learning

MPA marine predator algorithm

MSA multiscale attention

NLP natural language processing

NMS non-maximum suppression

NNs neural networks

NPUs neural processing units

PAN path aggregation network

R-CNN region-based convolutional neural network

R-CNN region-based convolutional neural network

ResNet residual neural network

RF Roboflow

RFO red fox optimization

RNNs recurrent neural networks

ROI region of interest

RPN region proposal network

RPN region proposal network

RT-DETR real-time detection transformer

SPPF spatial pyramid pooling-fast

SSD single-shot detector

SVM support vector machines

SVMs support vector machines

SW software

TPUs tensor processing units

UAVs unmanned aerial vehicles

VGG19 deep convolutional neural network by Visual Geometry Group at University of Oxford

ViT vision transformer

XAI explainable artificial intelligence

YOLO you only look once

Appendices

A1: **Diploma Thesis Assignment** (Zadání diplomové práce)

A2: **Source code in Jupyter Notebook** (XAI_GradCAM_YOLOv10_v5_0_0.ipynb)

UNIVERZITA HRADEC KRÁLOVÉ

Fakulta informatiky a managementu

Akademický rok: 2025/2026

ZADÁNÍ DIPLOMOVÉ PRÁCE

(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení: Bc. Richard Hlubuček
Osobní číslo: I2400694
Studijní program: N0613A140042 Aplikovaná informatika
Téma práce: Zpracování mikroskopických dat metodami strojového učení
Zadávající katedra: Katedra informačních technologií

Zásady pro vypracování

[1] Introduction
[2] Related Work and Motivation
[3] State-of-the-Art
[4] Methodology
[5] Results
[6] Discussion
[7] Conclusion and Future Work
Bibliography

Rozsah pracovní zprávy:
Rozsah grafických prací:
Forma zpracování diplomové práce: tištěná/elektronická
Jazyk zpracování: Angličtina

Seznam doporučené literatury:

- [1] Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep Learning. MIT Press. ISBN 9780262035613.
- [2] Alpaydm, E. (2014) Introduction to Machine Learning, third edition. MIT Press. ISBN 9780262028189.
- [3] Hutter, F., Kotthoff, L. and Vanschoren, J. (2019) Automated Machine Learning: Methods, Systems, Challenges. Springer Nature. ISBN 9783030053178 (print); 9783030053185 (eBook). DOI 10.1007/978-3-030-05318-5.
- [4] Russell, S.J. and Norvig, P. (2022) Artificial Intelligence: A Modern Approach, fourth edition. Pearson Education Limited. ISBN 9781292401133.
- [5] Kelleher, J.D., Mac Namee, B. and D'Arcy, A. (2020) Fundamentals of Machine Learning for Predictive Data Analytics, second edition: Algorithms, Worked Examples, and Case Studies. MIT Press. ISBN 9780262044691.
- [6] Jung, A. (2022) Machine Learning: The Basics. Machine Learning: Foundations, Methodologies, and Applications. Springer Nature Singapore Pte Ltd. ISBN 9789811681929. DOI 10.1007/978-981-16-8193-6.
- [7] Buduma, N. and Lacascio, N. (2017) Fundamentals of Deep Learning. O'Reilly Media. ISBN 9781491925614.
- [8] Han Veiga, M. and Ged, F.G. (2024) The Mathematics of Machine Learning: Lectures on Supervised Methods and Beyond. Walter de Gruyter. ISBN 9783111288475.
- [9] Kelleher, J.D. (2019) Deep Learning. MIT Press. ISBN 9780262537551.

Vedoucí diplomové práce: **doc. RNDr. Kamila Štekerová, Ph.D., MSc.**
Katedra informačních technologií

Oponent diplomové práce: **Ing. Martina Husáková, Ph.D.**
Katedra informačních technologií

Datum zadání diplomové práce: **7. října 2025**

prof. Ing. Mgr. Petra Marešová, Ph.D., MBA v.r.
děkanka

LS.

doc. Ing. Pavel Čech, Ph.D. v.r.
vedoucí katedry

V Hradci Králové dne 7. října 2025