UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Computer Science Curriculum

Aleksandr Makarov

Mastering the Unseen: Approaches to Hard-to-Detect Viral Cytopathic Effect

Master's Thesis (30 ECTS)

Supervisor: Dmytro Fishman, PhD

Tartu 2025

Mastering the Unseen:
Approaches to Hard-to-Detect Viral Cytopathic Effect

Abstract:
Viral infections pose persistent global health challenges, making rapid, accurate assessment of viral activity crucial for research and diagnostics. Cytopathic Effect (CPE), morphological changes in host cells upon viral infection, serves as a critical visual indicator of viral load, yet its manual microscopy based assessment is laborious and subjective. Furthermore, simple automated classification often fails to quantify infection severity and struggles with "hard-to-detect" cases. This work presents a comprehensive survey and performance evaluation of various computer vision techniques, ranging from image classification to weakly and strongly supervised segmentation with both classical and deep learning-based models, for the automated detection and localisation of CPE induced by xenotrophic murine leukemia virus (x-MuLV). Our analysis demonstrates that supervised segmentation techniques provide a significantly more robust pathway for viral load quantification than explainability-based classification methods, particularly when analysing images displaying subtle cellular alterations with low viral load. This automated methodology offers an efficient, objective, and scalable alternative to manual inspection, facilitating high-throughput analysis and deeper insights into infection dynamics. Following extensive data preparation, this work systematically compared existing computer vision methodologies, thereby identifying and validating best-performing approaches for consistent and quantitative Cytopathic Effect characterisation, which offers a powerful tool to accelerate drug discovery, advance fundamental viral research, and improve automated virological assays.

Keywords:
Biomedical Computer Vision, Deep Learning, Viral Cytopathic Effect, CPE

CERCS: T111 - Imaging, image processing; P176 - Artificial intelligence; B110 - Bioinformatics, medical informatics, biomathematics biometrics

Nähtamatu Taltsutamine: Lähenemisviisid Raskesti Avastatava Viiruse Tsütopaatilise Efekti Tuvastamiseks

Lühikokkuvõte: Viirusnakkused on maailma tervisele pidevaks väljakutseks, mistõttu on viiruse aktiivsuse kiire ja täpne hindamine väga oluline nii teadustöös kui diagnostikas. Tsütopaatiline efekt (CPE) – viiruse põhjustatud morfoloogilised muutused peremeesrakkudes – on tähtis visuaalne näitaja viiruskoormuse kohta, kuid mikroskoobiga käsitsi hindamine on aeg-nõudev ja subjektiivne. Lisaks ei suuda lihtsad automaatsed klassifikaatorid tihti nakkuse raskusastet täpselt mõõta ning jäävad hätta «rasketuvastatavate» juhtumitega. Siin uuringus anname põhjaliku ülevaate ja hindame eri arvutinägemise võtteid – alates pildiklassifikatsioonist kuni nõrgalt ja tugevalt juhendatud segmenteerimiseni, kasutades nii klassikalisi kui sügavõppe mudeleid – xenotroopse hiire leukeemiaviiruse (x-MuLV) poolt tekitatud CPE automaatseks tuvastamiseks ja lokaliseerimiseks. Analüüs näitab, et juhendatud segmenteerimismeetodid on viiruskoormuse kvantifitseerimisel märksa töökindlamad kui selgitavusel põhinevad klassifitseerijad, eriti kui piltidel on vaid õrnad rakumuutused ja madal viiruskoormus. See automatiseeritud lahendus pakub tõhusat, objektiivset ja skaleeritavat alternatiivi käsitsi vaatlemisele, võimaldades kõrge läbilaskevõimega analüüsi ning sügavamat arusaama infektsiooni dünaamikast. Pärast ulatuslikku andmete ettevalmistust võrreldi süstemaatiliselt olemasolevaid arvutinägemise meetodeid ning tuvastati ja kinnitati parimad lähenemised CPE järjepidevaks ning kvantitatiivseks iseloomustamiseks. Tulemuseks on võimas tööriist, mis kiirendab ravimite avastamist, toetab viroloogia alusuuringuid ja parandab automatiseeritud viroloogilisi analüüse.

Võtmesõnad:
Biomeditsiiniline arvutinägemine, süvaõpe, viiruslik tsütopaatiline efekt

CERCS: T111 - Pilditehnika; P176 - Tehisintellekt; B110 - Bioinformaatika, meditsiiniinformaatika, biomate maatika, biomeetrika

Contents

1 Introduction

1.1 Problem and Motivation

Viral infections represent persistent global health challenges, with certain viruses possessing the capacity to spread rapidly and cause widespread public health crises, as tragically highlighted by past pandemic [1]. In responding to such threats, the development and deployment of rapid and accurate diagnostic techniques are vital. Within virological research and diagnostics, cell cultures are indispensable tools for studying viral behaviour, screening antiviral compounds, and producing vaccines. A critical indicator of viral activity in these cultures is the Cytopathic Effect (CPE), which is a common symptom for many viruses. CPE refers to a spectrum of morphological changes in host cells caused by viral infection, including alterations in cell shape, detachment from the culture substrate, fusion into giant cells, or the appearance of intracellular inclusion bodies [2]. Detecting and characterising CPE is thus fundamental for confirming viral presence, understanding mechanisms of viral pathogenesis, and evaluating the efficacy of antiviral therapies [3].

Traditionally, the identification and assessment of CPE in cell cultures are performed manually by trained laboratory personnel observing samples under brightfiled microscopy. While established, this manual process is inherently labour-intensive, time-consuming, and susceptible to inter-observer variability, particularly when distinguishing subtle "hard-to-detect" CPE. This matter is further complicated since different viruses have unique CPE [2, 4]. In an era where high-throughput screening and rapid data turnaround are increasingly critical, these limitations can create significant bottlenecks in research and diagnostic workflows, underscoring the need for more efficient and objective solutions.

The field of biomedical computer vision, propelled by significant advancements in machine learning methodologies, particularly deep learning [5], and the growing availability of large-scale image datasets [6], offers transformative potential in this domain. Tasks previously deemed impractical or reliant on extensive manual expert intervention are now possible to automate. By leveraging these computational tools, biomedical image analysis can be significantly enhanced, allowing researchers and medical professionals to focus on more complex interpretative tasks, thereby improving overall efficiency and the reliability of outcomes in both research and clinical settings.

This research is motivated by the potential to significantly streamline viral CPE analysis workflows, reduce subjectivity, and provide more rapid, detailed insights for virological research.

1.2 Goals

This thesis addresses the pressing need for automated and quantitative analysis of viral CPE by developing a computer vision-based pipeline. The research specifically focuses on detecting and characterising CPE induced by xenotrophic murine leukemia virus

(x-MuLV), a model retrovirus, also used for the development of gene therapy [7].

The primary objectives of this study are:

1. To develop robust machine learning models capable of accurately detecting the presence of viral CPE, distinguishing between infected and uninfected cell populations from bright-field microscopy images.

2. To move beyond binary (presence/absence) classification by exploring and implementing segmentation approaches. These techniques aim to localise specific regions exhibiting CPE, thereby providing more granular information that can be used to quantify the extent or severity of viral infection within an image or well. This quantitative aspect is crucial for applications such as viral load estimation or detailed tracking of infection dynamics.

1.3 Contribution

This thesis makes a primary contribution through the comprehensive application and comparative evaluation of various computer vision methodologies, from classical algorithms to deep learning models, for the detection and, importantly, the quantification of xenotrophic murine leukemia virus-induced Cytopathic Effect. A distinctive aspect of this work is its systematic investigation of segmentation for CPE characterisation. While prior automated approaches to CPE detection have largely focused on image-level classification [4, 8, 9, 10], this research demonstrates that segmentation can yield granular, quantitative insights into the extent of viral infection. Consequently, this study establishes the feasibility of leveraging automated techniques for a more efficient, objective, and detailed assessment of viral load, offering substantial benefits over traditional manual methods, especially for challenging "hard-to-detect" CPE manifestations.

1.4 Thesis structure

The thesis is structured as follows:

1. **Background**: This section provides a comprehensive overview of the viral CPE problem and explores relevant computer vision techniques applied in the biomedical domain.

2. **Data and Methods**: This section details the datasets and methodologies employed in this study, including the specific approaches used for classification and segmentation.

3. **Results and Discussion**: This section presents the findings obtained from the application of the various methods and approaches described in the previous

section and interprets the results, assessing their significance and implications within the context of the research question and previous studies.

4. **Conclusion**: This section summarises the key findings of the study and outlines potential directions for future research.

2 Background

Computer vision has revolutionised biomedical imaging, particularly in classification [11], object detection [12], and segmentation [13], driven by large datasets and advancements in image processing and hardware [14]. This chapter focuses on the application of these techniques to viral CPE detection. It is vital for identifying viral presence, studying infection dynamics, and assessing antiviral treatments in virology. However, a lot of labour is devoted towards manual labelling of infected cell lines on brightfield microscopy images. This section introduces the significance of CPE and then explores the techniques that enable not only the detection but also the detailed segmentation-based quantification. These techniques range from traditional image processing to deep learning approaches, including Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). Given the challenges of limited annotations in biological data, we then examine Multiple Instance Learning for weakly labelled scenarios. Finally, we discuss Explainable AI techniques, which are crucial for interpreting and validating complex models.

2.1 Viral Cytopathic Effect

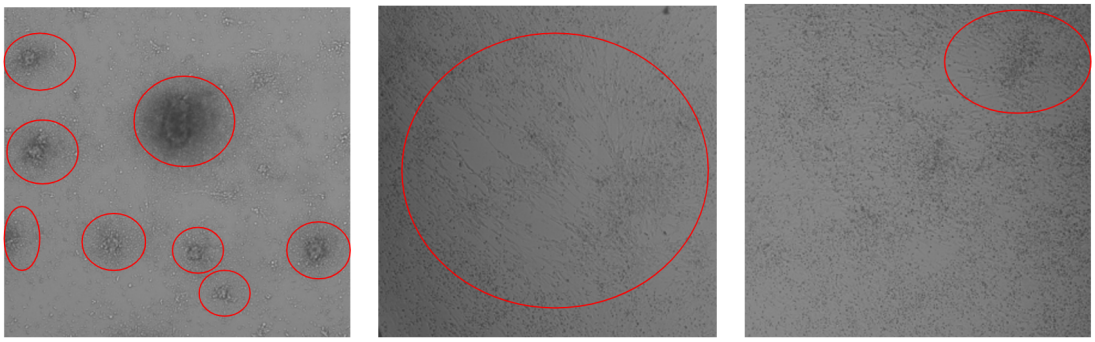


Figure 1. Examples of brightfiled microscopy images of infected and uninfected cells with strong cell death and aggregation (a), star-like morphology (b) and healthy cells, with possible signs of cytotoxicity in the top-right corner (c).

The study of viral infections relies heavily on cell cultures, where viruses are incubated to assess treatment efficacy and conduct routine analyses. CPE, visually detectable through bright-field microscopy, has long been a critical metric for identifying and quantifying viral activity [15]. However, the complexity of CPE lies in its diverse presentation (see Figure 1). Different viruses induce a spectrum of cellular changes, including cell death, lysis, morphological variations (rounding, enlargement, star-like formation), functional disruption, granulation, and detachment [2, 3]. This variability, coupled with the

fact that not all viruses exert CPE in cell cultures, defines the need for supplementary detection methods [16].

Many laboratory techniques are available for virus detection, with one of the most commonly employed being the Tissue Culture Infectious Dose ($TCID_{50}$) assay [17]. This method utilises cell stains, such as crystal violet, but typically requires several days to generate results. Another widely used approach involves conventional brightfield microscopy. However, this method is both time-consuming and labour-intensive due to the necessity for manual inspection and labelling. Additionally, the development of visible CPE can take between 1 and 14 days, depending on factors such as viral load and virus type [17], which might not be the optimal if real time detection is needed and more labor intense techniques should be used instead such as green fluorescent protein (GFP) labeling, monoclonal antibodies and ATeam probe [15].

To overcome the limitations of traditional methods, recent research has focused on developing automated and rapid computer vision methodologies for viral CPE detection [8]. Studies have demonstrated the feasibility of using deep learning for classifying CPE induced by various viruses [9, 10]. For instance, researchers successfully developed a computer vision model for influenza virus CPE detection [10]. Dodkins et al. achieved accurate CPE identification and discrimination across four major viral categories (DNA, RNA, enveloped, and non-enveloped) [9]. Another study showed a strong correlation (r^2 = 0.986) between predicted $TCID_{50}$ and ground truth for eight human-affecting viruses, further validated by activation maps that aligned with plaques and distinct CPE features [4]. More recently, similar methodologies have been applied to successfully detect animal viruses in two cell lines [8]. In a more advanced approach, spatial light-interference microscopy has been utilised to identify viral agents, including SARS-CoV-2 [18].

The specific focus of this thesis is xenotrophic murine leukemia virus (x-MuLV), a retrovirus known to induce cancer in murine hosts and capable of infecting other species. This virus was inoculated in the PG4 cell line of cat glial cells. This particular virus-cell line combination is known to exhibit a range of symptomatic expressions and is valuable for quantification assays [19]. The virus induces distinct CPE morphology, and automatic recognition has not been demonstrated for it in the literature. Apart from the virus, the novelty of this work is that related research used "label-free" approach [4, 8, 9, 10] and did not use segmentation masks. "Label-free" means that there were no precise annotations of the infected area, but only one class per image. This approach is discussed in the following sections, starting with classical computer vision and machine learning techniques, followed by their deep learning successors.

2.2 Classic Computer Vision Approaches

Computer vision, the field of enabling machines to interpret and understand visual data from the world, has evolved significantly over the decades [20]. Classical approaches to computer vision laid the foundations for modern techniques. These methods, which

operate without the use of neural networks, provide a foundation for understanding image processing and analysis, and remain relevant for specific applications [20]. In the context of viral CPE detection, classical computer vision offers valuable tools for initial analysis and feature extraction.

Among these techniques, thresholding is a fundamental method for image segmentation, effectively separating foreground from background based on pixel intensity. One particular algorithm, Otsu thresholding [21], automatically sets the threshold between 2 classes by maximising inter-class variance. This approach can be particularly useful for identifying regions of severe CPE, where significant cellular changes are evident. Edge detection algorithms, such as Canny [22] and Sobel [23], play a crucial role in delineating boundaries within images by detecting abrupt changes in intensity. These algorithms can aid in highlighting distinct CPE features or, when combined with blob detection, in identifying individual wells in cell culture plates. Feature descriptors methods, such as histogram of oriented gradients (HOG) [24], were found to be useful as feature extractors coupled with classical machine learning methods for viral CPE detection [4].

Classical computer vision offers several advantages, particularly in scenarios where computational resources are limited or data scarcity is a concern [20]. These methods are often straightforward to implement, requiring less computational power than deep learning models, making them suitable for real-time applications. Furthermore, classical techniques are inherently more interpretable, allowing researchers to understand and debug algorithms more readily and identify potential sources of bias. This transparency is crucial in biomedical applications where understanding the underlying processes is essential. Additionally, these methods can be effective with limited data, a significant advantage in situations where data collection is challenging or expensive, or where the task is simple, such as basic colour detection.

However, classical computer vision approaches are not without limitations. Their performance heavily relies on the quality of hand-crafted features, which require domain expertise and can be time-consuming to develop [14]. Moreover, these techniques can be sensitive to variations in input data, such as changes in illumination, occlusion, or noise. This sensitivity poses a significant challenge in biomedical imaging, where such variations are common and can significantly impact analysis accuracy. Therefore, while classical methods provide a valuable starting point, they may not be sufficient for the nuanced and complex analysis required for robust viral CPE detection in diverse experimental conditions.

2.3 Deep Learning Computer Vision Approaches

Deep learning, a subset of machine learning, has revolutionised computer vision and other domains, achieving state-of-the-art performance in areas ranging from robotics to biomedical applications [25]. The field's rapid advancement is attributed to increased dataset availability, methodological innovations, and hardware improvements. The 2012

AlexNet study [11] marked a turning point, enabling end-to-end learning at scale and eliminating the need for manual feature engineering [14]. This advancement was further facilitated by the release of large-scale datasets such as ImageNet and COCO [26, 27]. Computer vision, well-suited for deep learning, has benefited significantly from these advancements, addressing tasks like classification, object detection, segmentation, and generative tasks such as image-to-image regression and enhancement [28].

At its core, deep learning builds upon the fundamental Artificial Neural Network (ANN) model, the perceptron [29, 30]. A perceptron consists of artificial neurons that perform weighted sums of inputs with biases, adjusted through backpropagation based on a loss function [31, 32]. Non-linear activation functions are applied after each neuron computation, enabling the network to approximate complex functions. The power of deep learning arises from expanding the network's width (neurons per layer) and depth (number of layers), giving name to deep learning and allowing it to learn intricate patterns, as demonstrated in tasks like handwritten digit recognition [33]. However, for more complex tasks, more specific architectures are required, which will be discussed in the following chapters.

Supervised training, where input samples are paired with corresponding labels, is the primary learning paradigm in deep learning. In computer vision, these labels can range from image categories to object bounding boxes, segmentations, or descriptive text. For viral CPE detection, image classification, object detection, and segmentation are particularly relevant.

Deep learning offers several advantages over classical computer vision, including end-to-end learning, scalability, and adaptability. End-to-end learning eliminates the need for manual feature extraction, although annotation remains a challenge. Weakly supervised or self-supervised learning can mitigate this burden. Deep learning models can effectively process large datasets and generalise to unseen data, making them suitable for applications like medical imaging [25, 28]. Furthermore, these models can adapt to variations in input data, provided sufficient training data and appropriate augmentations are used. The double-descent effect, challenging the bias-variance tradeoff, is another notable phenomenon in deep learning [34]. However, deep learning's superiority is most evident in complex tasks with large datasets. While it integrates feature extraction and inference, hybrid approaches combining classical and deep learning techniques also exist [20].

2.3.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) have long been recognised for their exceptional ability to learn hierarchical visual patterns, drawing inspiration from the architecture of biological vision systems [32, 36]. However, their widespread adoption was initially hindered by the substantial computational demands of the backpropagation algorithm and the lack of efficient parallel processing on CPUs. This limitation was overcome

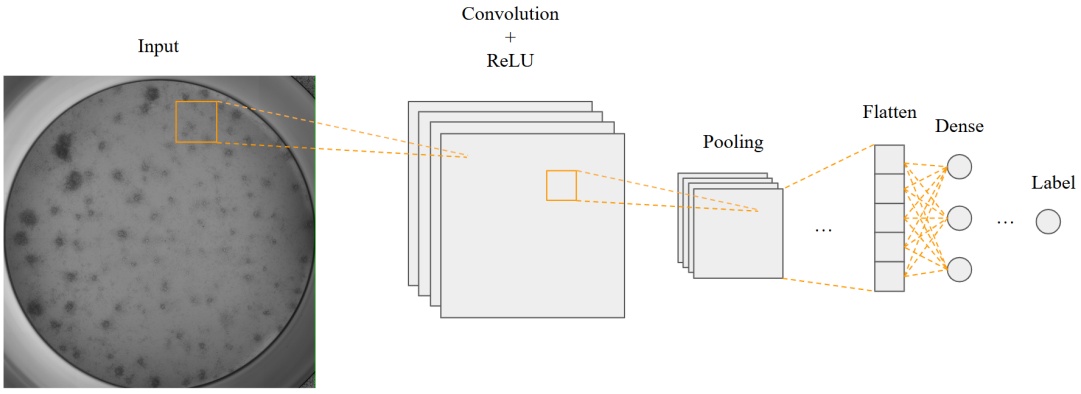


Figure 2. CNN structure [35]. Input signal is convolved through convolution kernels, activation functions, pooling(subsampling), by subsequent flattening and a feed-forward neural network.

with the advent of GPU acceleration, enabled by technologies like CUDA [37], which significantly accelerated the training of deep networks. This breakthrough propelled CNNs to victory in the ImageNet competition and, subsequently, surpassing human performance in certain visual recognition tasks [38, 39].

CNNs are a specialised class of deep learning models designed to process structured grid data, such as images. Their ability to automatically and adaptively learn spatial hierarchies of features has made them the cornerstone of numerous computer vision applications [5]. A typical CNN architecture comprises multiple layers, including convolutional layers, pooling layers, and fully connected layers. Convolutional layers apply learnable filters to input images, extracting relevant features. Pooling layers then downsample these feature maps, reducing spatial dimensions and computational complexity. Finally, fully connected layers perform the final classification based on the extracted features (Fig. 2). The network's weights and biases are adjusted through backpropagation, driven by a loss function, to optimise performance.

As CNN architectures evolved, researchers addressed various challenges to improve performance and efficiency. One significant obstacle was the vanishing and exploding gradients problem, which hindered the training of deeper networks [40]. This led to the development of innovative architectures, such as Inception [41], which focused on computational efficiency, and ResNet [42], which incorporated residual connections to mitigate gradient issues. More recently, models like EfficientNet [43] and EfficientNetV2 [44] have optimised architecture efficiency further, achieving state-of-the-art performance with improved efficiency at that time.

CNNs have demonstrated remarkable success across various computer vision tasks, including image classification, object detection, and semantic segmentation. In the

biomedical domain, CNN-based architectures have gained significant traction, with U-Net [13] becoming a widely adopted model for biomedical image segmentation. Its successors, such as nnU-Net [45], DC-UNet [46], and MultiResUNet [47], have further improved segmentation performance in medical applications.

For object detection and segmentation tasks, the Region-Based CNN (R-CNN) family [12], including Mask R-CNN [48], has achieved notable success. However, these models are computationally demanding, limiting their applicability in resource-constrained environments. To address this, the lightweight YOLO (You Only Look Once) family of models emerged [49], offering a balance between accuracy and efficiency, making them particularly suitable for real-time applications and deployment on low-end GPUs or CPU-only systems.

2.3.2 Transformers

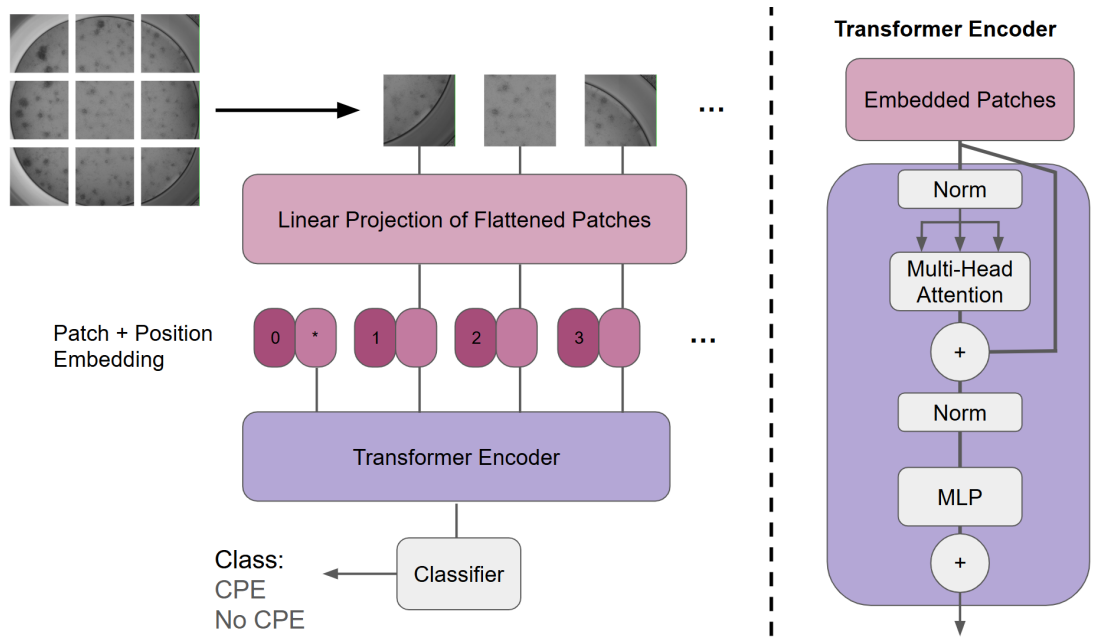


Figure 3. Diagram of ViT. Image patches are processed as input with position embedding, and then the transformer encoder uses a multi-layer perception as a classifier.

Originally developed for Natural Language Processing (NLP), transformers have revolutionised deep learning by utilising self-attention mechanisms to capture long-range dependencies in sequential data [50]. Their remarkable success in NLP prompted researchers to explore their applicability in computer vision, leading to the development of Vision Transformers (ViTs), which have emerged as a powerful alternative to Convolutional Neural Networks (CNNs) [51].

The first major adaptation of transformers to computer vision was the ViT [51], which demonstrated that self-attention-based architectures could achieve state-of-the-art performance in image classification when trained on extensive datasets. Unlike CNNs, which leverage hierarchical convolutional operations to extract spatial features, ViTs treat an image as a sequence of patches, mirroring how NLP transformers process words in a sentence.

ViTs first divide an input image into fixed-size patches, which are then flattened and linearly embedded into a lower-dimensional space, as depicted on Figure 3. Each patch embedding is combined with a positional encoding to retain spatial information, as transformers lack the intrinsic inductive biases of CNNs, such as locality and translation equivariance. These embedded patches are then processed by a standard transformer encoder, consisting of:

- Multi-Head Self-Attention (MHSA): Computes attention scores between all patches, allowing the model to capture both local and global relationships within the image.
- Feedforward Neural Networks (FFN): Fully connected layers with non-linear activations applied to each patch embedding.
- Layer Normalization and Skip Connections: Helps stabilize training and improve gradient flow, similar to ResNet-style architectures.

The final representation is processed by a classification head for image recognition tasks. This architecture enables ViTs to learn long-range dependencies more effectively than CNNs.

Following the success of ViT, several variants and improvements have been proposed to enhance efficiency and performance: DeiT (Data-efficient Image Transformer) [52] introduced knowledge distillation techniques to train ViTs with significantly less data, Swin transformer [53] incorporated hierarchical feature maps and shifted window attention to improve computational efficiency and adaptability to dense prediction tasks, PiT (Pooling-based Vision Transformer) [54] introduced pooling layers to reduce the number of tokens, mimicking the downsampling effect of CNNs.

ViTs offer several advantages over CNNs, particularly in their ability to capture global context and scale effectively with larger datasets. Their reliance on self-attention enables them to model long-range dependencies more efficiently, making them particularly well-suited for tasks such as image classification, object detection, and semantic segmentation. However, ViTs typically require extensive computational resources and large-scale datasets for effective training, as they lack the strong inductive biases of CNNs, which allow for more efficient learning with limited data.

2.3.3 Training approaches

Deep learning models can be trained using three primary approaches, each with distinct advantages and limitations:

1. **Training from Scratch**: This approach involves initialising the model with random weights and training it on a specific dataset. While it offers the highest flexibility for learning domain-specific features, it demands substantial amounts of labeled data and extensive computational resources. This method is typically employed when the target domain significantly differs from existing pre-training datasets.

2. **Transfer Learning**: This technique utilises a pre-trained model, often trained on a large dataset like ImageNet, as a feature extractor. The backbone layers of the model are frozen, and only the task-specific layers, such as a classification head, are trained on the new dataset. This approach is particularly advantageous when labelled data is scarce, as it significantly reduces training time while capitalising on knowledge acquired from previous datasets [55].

3. **Fine-Tuning**: Similar to transfer learning, fine-tuning involves using a pre-trained model. However, instead of freezing the backbone, the entire model or a portion of it is further trained on the new dataset using a lower learning rate. This allows the model to adapt more effectively to domain-specific characteristics while retaining valuable pre-trained knowledge [56]. Fine-tuning is most effective when the new dataset shares similarities with the original pre-training dataset.

While fine-tuning is highly effective for adapting models to similar tasks, it is less suitable for domains with significantly different data modalities or problem domains. In such scenarios, training new architectures from scratch is often necessary to learn task-specific representations. However, random weight initialisation can lead to suboptimal convergence and sensitivity to initialisation strategies [57, 58].

To address the challenges of training from scratch, particularly in low-data scenarios, recent research has explored improved weight initialisation techniques, architectural modifications, and self-supervised learning approaches. Moreover, hybrid strategies combining fine-tuning with novel training paradigms, such as contrastive learning and meta-learning, are being investigated to enhance model adaptation across diverse biomedical applications [25].

2.3.4 Multiple Instance Learning

Within machine learning, there is a specialised domain that deals with noisy, incomplete, or weakly labelled data, making it well-suited for weakly supervised learning. This approach falls within the broader category of semi-supervised and multiple-instance learning (MIL).

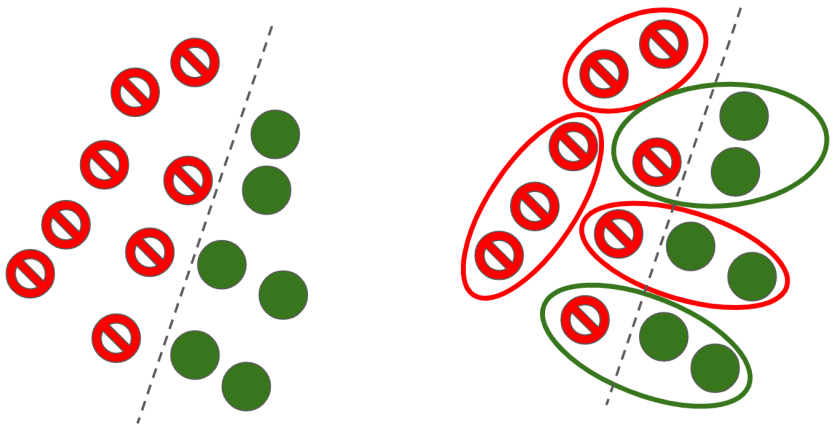


Figure 4. The visualisation of the difference between single instance learning (a) and multiple instance learning bags (b) for a binary classification problem.

In MIL, data is organised into bags of instances, where individual instances lack explicit labels, but a label is provided for the entire bag (Fig. 4). For example, in digital pathology, a whole-slide gigapixel image may be labelled as cancerous, but its individual patches (subregions) do not have specific annotations. The model must then learn to identify relevant patterns from weakly labelled data [6].

This type of weakly supervised learning is particularly valuable in histopathology, where vast amounts of imaging data are available but expert annotations are limited. As the number of medical professionals declines and AI-assisted diagnostics improve, MIL-based approaches are becoming increasingly popular in medical imaging and automated disease detection [6]. This approach is incredibly useful for the CPE detection problem since precise expert-labelled annotations for our task are unavailable.

2.4 Explainable AI in Computer Vision

Explainable AI (XAI) aims to make the decision-making process of AI models more understandable and transparent to humans. In the context of computer vision, XAI techniques help interpret the predictions made by deep learning models, providing insights into how the models arrive at their conclusions.

As computer vision models become more complex and integrated into critical applications, the need for explainability increases. XAI techniques enhance trust, accountability, and usability of AI systems by making their internal workings more comprehensible.

Various methods are used to achieve explainability in computer vision models. These include gradient-based methods, perturbation-based methods, and attention-based tech-

niques. Gradient-based methods, such as Grad-CAM [59], Grad-CAM++ [60], highlight the regions in the input image that contribute most to the model's prediction. Perturbation-based methods like LIME [61] or SHAP [62] involve modifying the input data and observing the changes in the model's output to understand its decision-making process. Attention-based techniques, specific to transformer architectures, provide insights into the model's focus areas within the input data, and are especially visible with Dino family models, where attention maps could be easily interpreted as segmentations [63, 64, 65, 66].

XAI techniques have been applied to various computer vision tasks, including medical imaging, autonomous driving, and quality control. In medical imaging, XAI helps clinicians understand the model's predictions and make informed decisions.

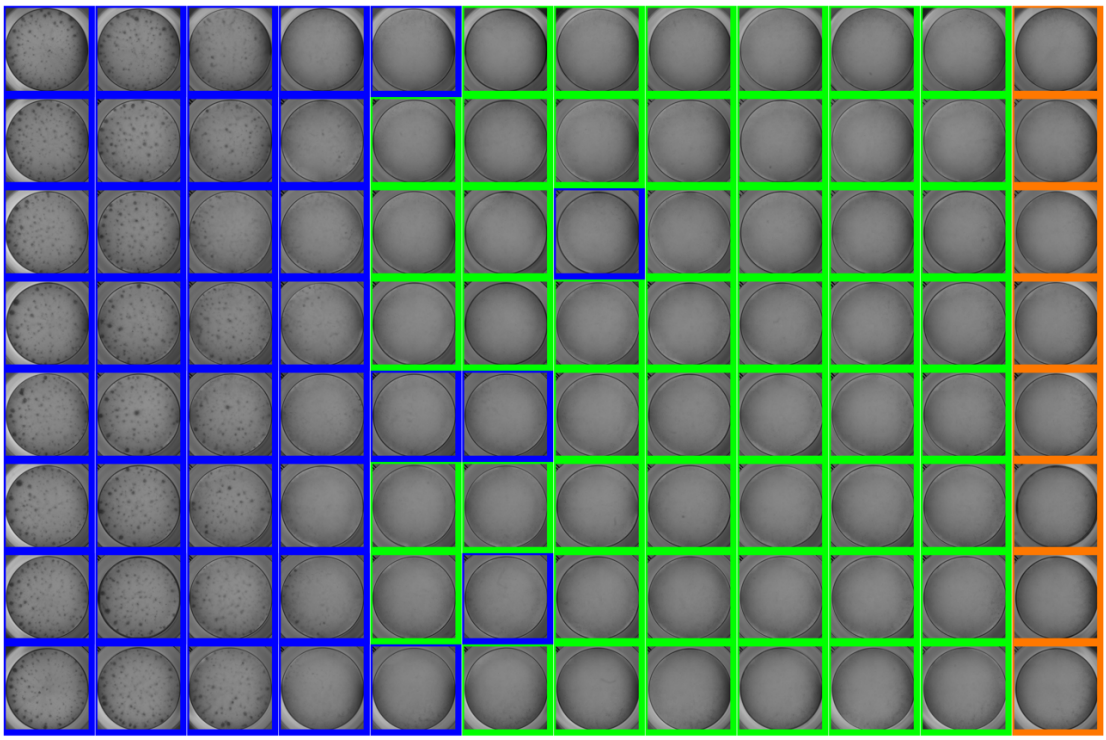


Figure 5. A 96-well plate illustrating the experimental layout, featuring a left-to-right decreasing virus concentration gradient. Well borders are color-coded according to visual inspection: blue for wells with evident CPE, green for virus-exposed wells without CPE, and orange for uninfected control wells. Most of CPE in wells beyond column 4, corresponding to lower virus titres, are considered "hard-to-detect" due to its subtle presentation.

3 Data and Methods

This section provides a comprehensive overview of the data and methodologies employed in this research. The primary objective was to develop robust models capable of accurately detecting both subtle and pronounced instances of viral CPE, while maintaining overall high performance. To achieve this, a range of techniques was investigated and evaluated, focusing on selecting those most suitable within the given constraints. This section begins with a detailed description of the dataset, followed by an exploration of the experimental approaches utilised for image classification, Multiple Instance Learning, and segmentation.

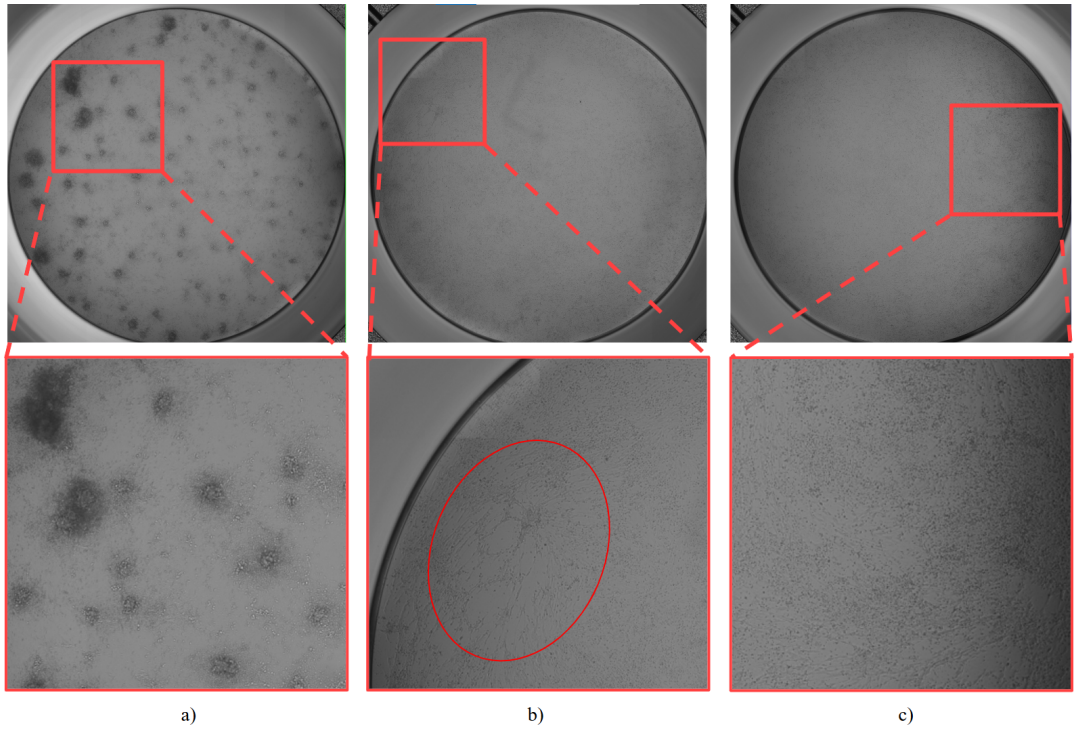


Figure 6. Representative bright-field microscopy images illustrating different manifestations of viral CPE compared with uninfected control cells. Zoomed-in patches are shown for detailed inspection. (a) Easy CPE cases are characterised by distinct and widespread cellular alterations such as cell rounding and aggregation(dark clusters). (b) A hard CPE case, displaying subtle and less obvious morphological changes indicative of viral infection; a representative area is highlighted by a red circle in the zoomed patch. (c) Uninfected control cells exhibited a healthy, confluent monolayer with normal morphology.

3.1 Dataset

The source dataset comprised 432 bright-field microscopy images of virally infected and uninfected cells, each with a resolution of 7546×7546 pixels and an 8-bit depth. These images, all annotated with image-level labels indicating the presence or absence of CPE, depict a single cell line infected with the same virus type.

This dataset originated from six 96-well plates. For this study, three plates were selected for detailed analysis based on their relative homogeneity in viral concentration gradients and consistent image stitching quality, yielding a total of 252 images. The remaining three plates were excluded due to significant out-of-distribution heterogeneity; this heterogeneity was primarily attributed to inconsistencies in viral concentrations and the presence of prominent image stitching artefacts (examples of such artefacts are

Table 1. Summary of the annotated image dataset, detailing the number of images classified into distinct categories based on visual assessment: "Easy CPE" (distinct viral effects), "Hard CPE" (subtle viral effects), and "No CPE" (uninfected/control).

Plate	Easy CPE	Hard CPE	no CPE	Total
H	27	11	58	96
L	31	11	54	96
X	24	4	54	60
Total	82	26	166	252

provided in supplementary Figure A.1. An overview of the dataset outline on 96-well plate is presented in Figure 5.

Viral CPE refers to a range of morphological changes in host cells resulting from viral infection, which can include cell rounding, shrinkage, detachment from the substrate, syncytia formation (fusion of multiple cells), and the appearance of inclusion bodies. In this context, "easy" CPE cases are those where these morphological changes are pronounced and widespread, often recognisable with minimal magnification. In contrast, "hard" CPE cases exhibit more subtle or focal manifestations of these viral-induced cellular alterations, requiring more detailed examination for reliable identification when compared to uninfected cells (refer to Figure 6 for visual examples).

The selected 252 reliable image subset consists of 82 images exhibiting easily visually discernible CPE ("easy CPE"), 26 images presenting challenging, subtle CPE cases ("hard CPE"), and 166 images without any visible CPE ("no CPE"). A detailed summary of the data distribution per plate is available in Table 1. Despite selecting the best plates without clearly distinguishable artefacts, some are still present. Average per pixel image intensity is visualised in supplementary Figure A.2 and clearly shows stitches and brightness gradient.

The annotation process for this dataset was iterative and involved several refinement stages. Initial segmentation predictions for both easy and hard CPE cases were provided by a partner company using their proprietary algorithms. However, visual inspection revealed these initial predictions to be insufficiently accurate for training robust models. Consequently, a manual correction and re-annotation workflow was implemented using Label Studio [67], augmented by the Segment Anything Model (SAM) [68].

This refinement process involved:

1. Visual inspection of images to identify regions exhibiting CPE-like morphology that were either incorrectly segmented or entirely missed by the initial proprietary algorithms.

2. Manual annotation of these identified regions with bounding boxes.

3. Utilization of these bounding boxes as prompts for SAM to generate initial segmentation masks.

4. Manual correction and refinement of the SAM-generated segmentation masks to ensure they accurately delineated the humanly discernible CPE structures.

Furthermore, during the development and testing of our own segmentation models, instances of previously overlooked CPE were identified in images initially labelled as "no CPE". These findings prompted a further round of re-annotation, guided by retrained model predictions, to enhance the overall accuracy and completeness of the ground truth dataset.

3.2 Applied Methods for CPE Detection

This subsection outlines the experimental approaches employed for CPE detection. Initially, with only image class labels available, image classification and multiple instance learning were explored. As algorithmic predictions and manual annotations became accessible, the methodologies were subsequently enhanced with segmentation and instance segmentation techniques.

3.2.1 CPE Classification

For the automated detection and classification of viral CPE in images (categorising them as infected or not infected), an image classification approach was adopted. Utilising the available image annotations, this aimed to establish a baseline using both classical feature descriptors and deep learning models. The simplicity and efficiency of this approach made it a suitable initial step in the analysis. CPE image classification was explored using classical computer vision techniques combined with classical machine learning algorithms, and the results were compared against those obtained using deep learning models such as ResNet, EfficientNet, ViT, and Swin Transformer. Various training strategies were implemented, including training from scratch, fine-tuning, and transfer learning.

To establish a baseline with classical methods, HOG features were extracted from the images after resizing them to 1600×1600 pixels. To enhance the signal-to-noise ratio and focus on relevant information, images were cropped to exclude areas outside the well boundaries. The parameters for the hog function from scikit-image [69] were set as follows: orientations=8, pixels_per_block=(32, 32), and blocks_per_cell=(2, 2). This configuration resulted in a flattened 59008-length feature vector per image. The resulting HOG features were then normalised to a consistent scale by subtracting the mean and dividing by the standard deviation. Principal Component Analysis (PCA) was applied to reduce dimensionality for visual inspection of possible heterogeneity and clustering. Scaled HOG features were subsequently used to train classical machine learning models,

including Logistic Regression, Nearest Neighbors, Support Vector Machines (SVMs), Random Forest, and Extreme Gradient Boosting (XGBoost). To ensure robust evaluation, an 8-fold cross-validation strategy, grouped by the physical row of 96-well plates, was employed. Normalisation was performed independently for the training and validation splits within each fold.

For deep learning-based classification, models from the ResNet and EfficientNet families of classical CNNs were utilised. Transformer-based models, ViT and Swin Transformer, were also included in the experiments. These models were trained using images scaled to 1024×1024 pixels and their corresponding labels. To mitigate overfitting, checkpoints corresponding to the best validation accuracy were saved, and an early stopping patience of 10 epochs, based on validation loss, was set. This ensured that training was terminated if no improvement in validation accuracy was observed for 10 consecutive epochs. For model evaluation on the test data, the weights from the best validation checkpoint were restored. Several augmentation strategies were employed to increase data variability and improve model robustness, such as random cropping and random 360-degree rotation (due to the circular shape of the wells), as well as brightness and contrast alterations.

3.2.2 Multiple Instance Learning

Multiple Instance Learning provides a suitable computational framework for addressing weakly supervised learning problems inherent in biological imaging, such as viral CPE detection. In typical CPE assays, annotations are assigned at the image level (e.g., "infected" or "uninfected"), which constitutes weak supervision because the underlying cellular landscape is often heterogeneous. Images labelled as "infected" may contain a mixture of healthy cells and cells exhibiting CPE, particularly at early time points or low viral concentrations. This discrepancy between the coarse annotation level (the "bag") and the fine-grained, spatially varying biological reality (the "instances" or regions within the image) necessitates an MIL approach.

In this study, we adopted an MIL strategy leveraging deep learning features. Image patches were generated by partitioning the area within each image boundary into a 20×20 grid. Each resulting patch was downscaled to 224×224 pixels to serve as input for feature extraction. A ResNet50 model, pre-trained on ImageNet, was utilised as a convolutional neural network (CNN) backbone to generate high-dimensional feature embeddings for each patch, capturing salient visual characteristics. These patch-level embeddings (instances) collectively form the representation of the image (bag).

The instance embeddings for each bag were subsequently processed by an attention-based deep MIL model, drawing inspiration from the architecture proposed by Ilse et al. [70]. This model employs an attention mechanism to learn the relative importance of each instance within the bag for the final classification task. The attention weights allow the model to focus on patches most indicative of CPE. Weighted instance features are

then aggregated to form a bag-level representation, which is subsequently used to classify the entire image. Model parameters, including fine-tuning of the ResNet50 backbone and the MIL aggregator, were optimised via backpropagation based on the image-level labels.

3.3 Applied Methods for CPE Localisation

3.3.1 CPE Localisation via Explainable AI of Classification

Both classical and deep learning methods offer tools to provide insights into model decision-making. Generally, simpler models (e.g., linear models, decision trees) are more inherently interpretable, while more advanced models like Random Forest and deep learning models need specific explainability techniques.

The interpretability of linear models is straightforward, achieved by examining the model coefficients. Each coefficient directly indicates feature importance, assuming features were normalized during training. The linear feature importance derived from HOG can then be mapped back to the pixels from which the features originated. However, a limitation exists: each HOG feature has a $2\times2\times8$ dimensionality and cannot be directly visualised in 2D. To address this, the features were mapped back to the original image space by aggregating the average coefficient value for each pixel. To tailor the explanation to a specific image, the HOG feature vector of that image (after normalisation) can be multiplied by the model coefficients to provide insights into the model's decision-making process for that particular instance.

Explainability with deep learning models is more challenging due to the inherently interdependent nature of their internal representations. Among the various methods that work relatively well for both CNN-based and transformer-based vision models are Grad-CAM (Gradient-weighted Class Activation Mapping), which weights 2D activations by the average gradient, and Grad-CAM++, an extension of the former method that utilises second-order gradients. The resulting important areas identified by these methods can then be segmented to highlight regions indicative of CPE.

3.3.2 CPE Localisation through Fully Supervised Segmentation

To achieve precise localisation of areas exhibiting viral Cytopathic Effect within the microscopy images, fully supervised segmentation techniques were used. This involved training models on images with corresponding pixel-level annotation masks, which were initially provided by a partner company and subsequently refined and augmented through the iterative process described in the Dataset section. Two main deep learning architectural families were investigated for this purpose: U-Net for semantic segmentation and YOLO-family models for instance segmentation.

For pixel-wise classification of CPE regions (semantic segmentation), the U-Net architecture was selected due to its proven efficacy in biomedical image segmentation [13, 45]. The U-Net is characterised by its symmetric encoder-decoder structure:

- The encoder (contracting path) consists of a series of convolutional and max-pooling layers that progressively downsample the input image, capturing hierarchical contextual features at increasing levels of abstraction.
- The decoder (expansive path) then symmetrically upsamples these feature maps using up-convolutions (or transposed convolutions), while concatenating them with high-resolution features from corresponding stages in the encoder via skip connections. These skip connections are crucial as they allow the network to combine precise low-level detail (important for accurate boundary localisation) with abstract high-level semantic information (important for class identification), enabling accurate delineation of segmentation boundaries.

To enhance feature extraction capabilities and potentially accelerate convergence, the standard encoder of the U-Net was replaced with a pre-trained EfficientNet-B0 model. EfficientNet-B0, known for its excellent balance of high accuracy and computational efficiency [43]. Its weights, pre-trained on the ImageNet dataset, provided a rich set of general visual features that were subsequently fine-tuned for the specific task of CPE segmentation. The output feature maps from the various stages of the EfficientNet-B0 encoder were then fed into the U-Net's decoder, which performed the upsampling and produced the final pixel-level segmentation mask indicating CPE-affected areas.

For the distinct task of not only segmenting CPE but also identifying individual instances of CPE regions (e.g., separate cell clumps or affected areas), models from the YOLO family were selected. YOLO models are renowned for their efficiency and strong performance in object detection and instance segmentation. We utilised YOLO versions that natively support instance segmentation, providing both bounding box coordinates and pixel-level masks for each detected CPE instance. In the training class weights were introduced on the instance level.

Our investigation included YOLOv8 [49], a widely adopted iteration known for significant architectural advancements over its predecessors. Key features of YOLOv8 include:

- Backbone: An improved backbone based on Cross Stage Partial (CSP) principles, notably utilizing C2f modules. These modules, evolving from the C3 modules of earlier versions like YOLOv5, efficiently enhance feature extraction capabilities and improve gradient flow through the network by effectively combining high-level features with contextual information.

- Neck: The neck architecture typically integrates a Path Aggregation Network (PANet) combined with Feature Pyramid Network (FPN) structures. This design facilitates robust multi-scale feature fusion by effectively merging features from different backbone stages – the FPN creates a top-down pathway to propagate strong semantic features, while PANet adds a bottom-up pathway for enhanced localisation information.

- Head: Significantly, YOLOv8 employs an anchor-free detection head. This approach directly predicts object properties like centres and dimensions rather than relying on offsets to predefined anchor boxes, which simplifies the output stage and can improve generalisation across diverse object shapes and sizes. Furthermore, YOLOv8 often utilises a decoupled head, separating the classification, bounding box regression, and (for segmentation models) mask coefficient prediction tasks into different branches, which can lead to improved accuracy for each task. These design choices contribute to its well-regarded balance of speed and accuracy.

We also explored YOLOv11, which is designed to introduce a more efficient architecture aimed at enhancing small object detection and improving overall accuracy while maintaining the high inference speed characteristic of the YOLO family. The architecture of YOLOv11 optimises both speed and accuracy by building on advancements from earlier YOLO versions. The main architectural innovations in YOLOv11 revolve around:

- C3K2 Blocks: At the heart of YOLOv11's backbone are C3K2 blocks, an evolution of the CSP bottleneck concept. The C3K2 block optimizes information flow by splitting feature maps and applying a series of smaller 3x3 kernel convolutions, which are computationally more efficient than larger kernels while retaining the model's ability to capture essential image features at different network stages. The SiLU (Sigmoid Linear Unit) activation function is commonly used within these blocks.

- SPFF (Spatial Pyramid Pooling Fast) Module: YOLOv11 retains the SPFF module, designed to pool features from different regions of an image at varying scales. This improves the network's ability to capture contextual information and detect objects of different sizes, especially beneficial for small objects which have historically been a challenge for YOLO models.

- C2PSA (Cross Stage Partial with Spatial Attention) Block: One of the significant innovations in YOLOv11 is the C2PSA block. This module introduces attention mechanisms that allow the model to focus more effectively on important regions within an image, such as smaller or partially occluded objects, by emphasising spatial relevance in the feature maps, thereby improving detection and segmentation precision.

- Multi-Scale Prediction Head: Similar to earlier YOLO versions, YOLOv11 uses a multi-scale prediction head to detect objects (and produce masks) at different sizes. The head typically outputs predictions for three different scales (e.g., corresponding to low, medium, and high-resolution feature maps) generated by the backbone and neck.

3.3.3 Iterative Training and Refinement

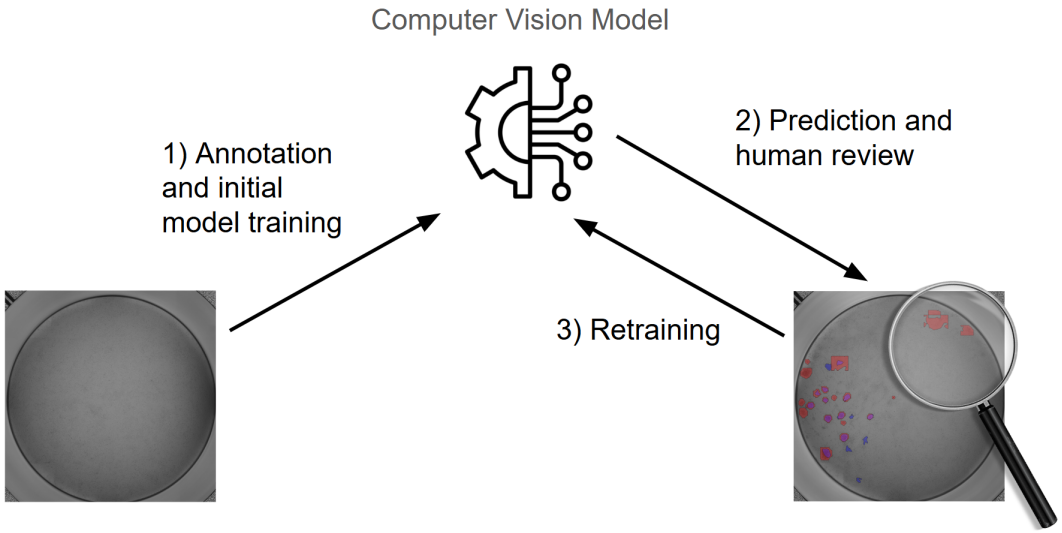


Figure 7. The flow diagram shows the summary of segmentation approach improvements.

An iterative refinement strategy was implemented to continually improve the model's detection and segmentation capabilities, particularly for challenging or subtly presented CPE cases. This process involved a cyclical workflow:

1. Initial Training: A baseline YOLO model was trained using the initially available set of manually annotated images.

2. Prediction & Review: The trained model was applied to predict CPE regions on a larger pool of images, including unannotated or partially annotated images.

3. Correction & Enrichment: These model predictions were manually reviewed. Corrections were made to inaccurate predictions (false positives, false negatives, inaccurate masks/boxes), and importantly, new annotations were added, with a focus on identifying previously missed difficult-to-detect ("hard") CPE instances.

4. Retraining: The enriched dataset, containing the original annotations plus the newly verified and corrected data, was used to retrain the YOLO model.

This human-in-the-loop approach facilitated annotation effort and progressive enhancement of the model's robustness as depicted in Figure 7. The training process evolved: initially, model development prioritised the accurate detection and segmentation of "hard" CPE cases, as these represented the primary diagnostic challenge due to their often subtle morphological changes. As the dataset expanded through iterative refinement cycles and incorporated more examples, including distinct ("easy") CPE morphologies, the model was subsequently trained for multi-class detection and segmentation, enabling it to differentiate between "hard" and "easy" CPE categories within the same framework.

3.4 Addressing Class Imbalance with Loss Weighting

A significant challenge during training was the inherent class imbalance within the annotated dataset. Instances corresponding to "hard" CPE were substantially less frequent than instances of "easy" CPE or background regions. Training directly on such imbalanced data typically results in models biased towards the majority classes, leading to poor sensitivity for the underrepresented "hard" CPE class.

To counteract this bias, a class weighting scheme was integrated into the model's loss function during training. This technique assigns different weights to the loss calculated for each class, effectively increasing the penalty for misclassifying samples from minority classes. The weight (w_j) for a given class (j) was determined using an inverse frequency calculation:

$$w_j = \frac{N}{C \times N_j}$$

Where:

- w_j represents the calculated weight assigned to class (j).
- N is the total number of annotated instances (e.g., segmented CPE regions) across all classes in the training dataset.
- C denotes the total number of distinct classes (e.g., background, easy CPE, hard CPE).
- N_j signifies the total number of annotated instances belonging specifically to class j.

By applying these weights, the contribution of the rarer "hard" CPE class to the overall training loss was amplified, compelling the model to learn its distinguishing features more effectively despite its limited sample size.

3.5 Evaluation

The performance of the developed models was assessed through a multi-faceted evaluation strategy operating at distinct levels, reflecting both the primary classification goal and the underlying instance segmentation task.

First, image classification accuracy was evaluated. Balanced accuracy was used from sklearn package, which is average recall obtained on each class. This represents the most critical benchmark from a practical standpoint, as the ground truth labels assigned to entire image (e.g., "infected" vs. "uninfected") are the most reliable annotations available in our dataset. Standard classification metrics, including f1 score, precision, recall, were calculated on a cross-validation. Cross validation was performed in a grouped manner by leaving out each row, resulting in 8 folds.

Second, segmentation performance was analyzed. The performance metric in this case was IoU. Segmentation models were not assessed for well classification purposes due to a high false-positive rate at uninfected wells.

Third, instance-level segmentation performance was analysed to evaluate the models' capability to accurately localise and/or segment individual regions exhibiting CPE. Annotating the same image twice by hand resulted in IoU < 0.7. Thus, instance-level metrics such as precision and recall were recorded based on an IoU 0.5 threshold. Crucially, given the focus on identifying different CPE morphologies, instance-level results were further stratified and reported separately for "easy" and "hard" CPE categories. This allowed for a granular understanding of model performance across varying levels of detection difficulty and validated the effectiveness of strategies employed to handle challenging cases, such as the class balancing applied during YOLO training.

3.6 Manuscript Refinement and Language Editing

To ensure clarity and accuracy in this thesis, several writing enhancement tools were used. The built-in grammar and spell-checking functionalities of Overleaf were employed for initial error correction. Additionally, Grammarly (Chrome extension) provided further assistance in refining grammar, spelling, and sentence structure for improved readability. For stylistic enhancements, some sections were reviewed and refined with the assistance of Large Language Models (LLMs), specifically Gemini 2.5 Pro and Flash. It is important to emphasise that these LLMS were used exclusively for rephrasing and improving the readability and academic tone of existing content, not for generating original text.

4 Results and Discussion

This section presents and discusses the experimental outcomes from the application of various methodologies to detect and characterise viral Cytopathic Effect. The findings are structured in the following order: we begin with an exploration of classical unsupervised analysis techniques, followed by supervised image classification using both classical machine learning algorithms and deep learning architectures. The interpretation of deep learning models is further enhanced through XAI techniques. Subsequently, the results from weakly-supervised MIL are presented. The section will then transition to evaluate approaches for CPE segmentation, encompassing both classical image processing and deep learning supervised and unsupervised strategies, finalising with an analysis of advanced instance segmentation performance using YOLO-family models.

4.1 Image classification

Our investigation commenced with image classification as a foundational step. The motivation was to first assess the task difficulty of Cytopathic Effect detection at an image level — a task considered easier as an initial approach compared to the direct segmentation of specific CPE regions, which would be required for detailed segmentation. This fully supervised strategy, where each microscopy image is associated with a single overall label (e.g., "CPE present" or "No CPE"), aimed to establish baseline performance levels for CPE detection and to compare the efficacy of classical machine learning techniques against modern deep learning models before progressing to more granular, localization-focused analyses.

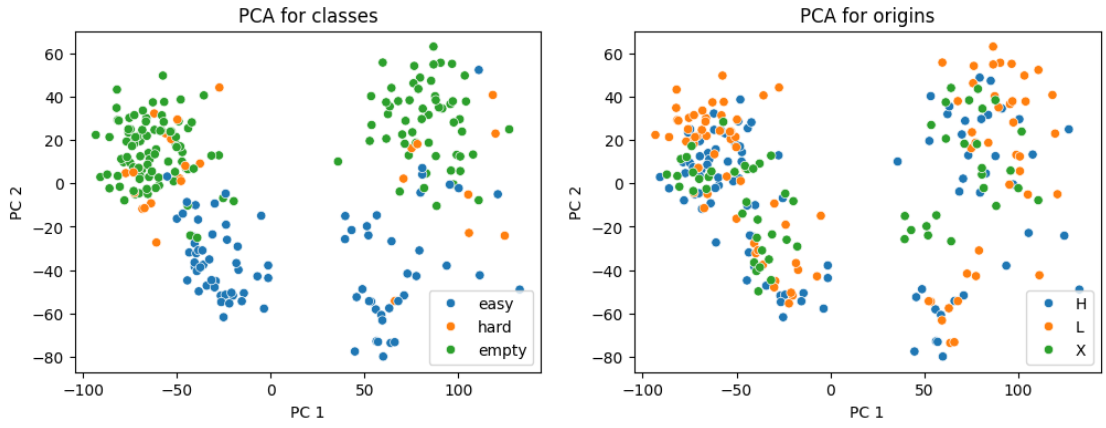


Figure 8. First 2 principal components of HOG features for all images with colour coding for different CPE types (a) and 96-well plates (b). H, L, X depict different plates.

4.1.1 Classical CPE Detection Approach

Classical machine learning methods for image classification traditionally depend on feature descriptors to represent image content. For this study, histogram of oriented gradients [24], as described in more details in Sections 2.2 and 3.2.1, was selected as the primary feature descriptor, valued for its robustness in capturing edge and shape information, which is important to identifying morphological changes indicative of CPE. HOG features were employed in both unsupervised and supervised workflows.

Initially, extracted HOG features from all images were standardised and subjected to Principal Component Analysis (PCA) for dimensionality reduction and exploratory data analysis. Inspection of the first two principal components aimed to reveal possible bias sources, such as clustering corresponding to CPE-infected versus uninfected wells, or distinctions between different experimental 96-well. As illustrated in Figure 8, this analysis revealed two distinct clusters in the PCA projection. However, these clusters did not directly align with the physical plate origins or pattern of well positions within the plates.

Another observation from the PCA plot was an apparent gradient across the data points, roughly corresponding to progression from easy to hard CPE to uninfected images. This gradient suggested that HOG features capture some discriminative information, indicating a moderate potential for successful classification using classical machine learning algorithms. The lack of distinct clustering for samples from different plates when viewed globally suggested that plate-specific features were not dominant, suggesting a moderate opportunity for models to generalise across different plates.

Table 2. Performance metrics for classical machine learning models trained on HOG features for Cytopathic Effect classification.

				Sensitivity		
Model Type	F1	Accuracy	Specificity	All	Easy	Hard
LogisticRegression	**0.84**	**0.86**	0.87	**0.84**	**0.93**	**0.58**
SVC	0.81	0.83	**0.94**	0.73	0.84	0.38
RandomForest	0.74	0.78	**0.94**	0.63	0.67	0.5
XGBoost	0.73	0.78	0.9	0.66	0.74	0.38
KNeighbors	0	0.5	1	0	0	0

Following unsupervised exploration, the normalised HOG features were utilised to train various supervised classical machine learning models to predict the CPE status of the corresponding images. A comprehensive comparison of these models, including Logistic Regression, Support Vector Classifier (SVC), Random Forest, XGBoost, and K-Nearest Neighbors, is presented in Table 2.

The results clearly indicate that Logistic Regression demonstrated the best overall

classification performance among the classical techniques tested. It achieved the highest F1-score (0.84) and accuracy (0.86). Notably, Logistic Regression also exhibited the strongest sensitivity, correctly identifying 93% of "easy CPE" cases and a leading 58% of the more challenging "hard CPE" cases. While its specificity was a robust 0.87, other models such as SVC and Random Forest achieved even higher specificity (0.94), indicating a very low rate of misclassifying uninfected samples. However, this came at the cost of lower sensitivity, particularly for "hard CPE" cases, where SVC and Random Forest only achieved sensitivities of 0.38 and 0.50, respectively. XGBoost's performance was comparable to Random Forest but did not surpass the leading models.

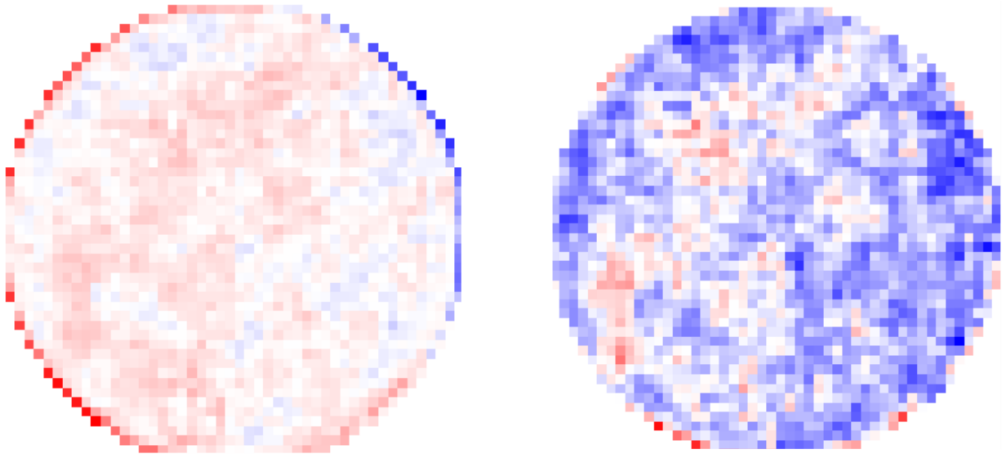


Figure 9. Visualisation of the average value of linear model coefficient per HOG descriptor cell. Higher value red, lower blue. a) represents initial setup with uncropped image with shortcut learning, which indicates that features at the left of the image add weight towards predicting CPE class, while higher value on the right results in higher weight for CPE absent class b) shows the same visualization but after the edge area was cropped much further and no significant gradient is present for edges.

Given its superior performance, the trained linear model was further analysed to gain insights into which HOG features were most influential in its predictions. The average coefficient values assigned to each HOG feature descriptor across cross-validation folds were examined. This explainability step proved crucial, as it led to the discovery of a shortcut learning phenomenon, depicted in Figure 9a. The model was found to be using artifactual gradients or features near the edges of images, rather than relying solely on cellular morphology. To mitigate this, a systematic image cropping was done, iteratively adjusting the crop until no significant feature importance was attributed to the extreme edges (Figure 9b). This corrected image preprocessing (cropping) was subsequently

adopted as a standard step for all downstream deep learning experiments to prevent similar shortcut learning issues. This was also taken into account for PCA plots.

4.1.2 Deep Learning CPE Detection Approach

Moving beyond classical techniques, various deep learning architectures and training strategies were explored for the CPE detection task. This included evaluating different model families, assessing the impact of fine-tuning pre-trained models versus transfer learning (with a frozen backbone) or training models entirely from scratch, and investigating the effects of class balancing.

Table 3. Performance comparison of different training strategies for the ResNet50 model in viral Cytopathic Effect image classification: fine-tuning pretrained ImageNet weights, training from scratch with random weights, and transfer learning with a frozen ImageNet pretrained backbone.

				Sensitivity		
Model&Weights	F1	Accuracy	Specificity	All	Easy	Hard
ResNet50 ImageNet weights	**0.89**	**0.9**	**0.99**	**0.81**	**0.99**	**0.27**
ResNet50 random weights	0.8	0.83	0.91	0.74	0.91	0.19
ResNet50 ImageNet weights frozen backbone	0.67	0.75	1	0.5	0.66	0

Initial experiments were conducted using the ResNet50 architecture, a widely adopted baseline for numerous computer vision tasks, to determine the optimal training strategy. The objective was to compare the efficacy of full fine-tuning of ImageNet pretrained weights, using the pretrained model as a fixed feature extractor (transfer learning with a frozen backbone), and training the model with randomly initialised weights (from scratch). The findings, summarized in Table 3, clearly indicate that fine-tuning the entire network from ImageNet pretrained weights yielded the best results among these ResNet50 configurations, achieving an F1-score of 0.89, an accuracy of 0.90, and perfect sensitivity (0.99) for "easy" CPE cases, though with a lower sensitivity of 0.27 for "hard" CPE cases.

Comparing this top ResNet50 performance to the best classical model (Logistic Regression with HOG features, which had an F1-score of 0.84), the fine-tuned ResNet50 demonstrated a notable improvement in overall F1-score and accuracy. However, it is crucial to highlight that for the challenging "hard" CPE cases, the classical Logistic Regression model showed superior sensitivity (0.58) compared to the fine-tuned

ResNet50 (0.27) at this stage. This suggests that while deep learning showed potential for higher overall performance, detecting subtle CPE remained a significant challenge where simpler models sometimes held an advantage, or this could be a potential shortcut indicator for classic methods, since no data augmentations were applied there.

The general success of fine-tuning is consistent with established practices [56, 55], as it allows the model to use the features learned from the vast ImageNet dataset while adapting them to the specifics of the CPE microscopy domain, often leading to faster convergence and superior performance. In contrast, using ResNet50 as a fixed feature extractor (transfer learning with a frozen backbone) resulted in considerably poorer performance, with an F1-score of only 0.67 and a complete failure to detect "hard" CPE cases (sensitivity of 0.00). This suggests that the high-level abstract features learned from ImageNet were not directly transferable or optimal for representing the distinct morphological features of viral CPE in cell lines without further adaptation, highlighting a significant domain shift. Training from randomly initialised weights, conversely, achieved the second-best performance among these three strategies, with an F1-score of 0.80 and a sensitivity of 0.19 for "hard CPE" cases, confirming the benefit of either learning domain-specific features from scratch or effectively adapting prelearned general features via fine-tuning.

Table 4. Comparative performance of various deep learning models for image classification-based detection of viral Cytopathic Effect.

				Sensitivity		
Model	F1	Accuracy	Specificity	All	Easy	Hard
EfficientNet-B0	**0.96**	**0.97**	0.99	**0.94**	**1**	**0.77**
ResNet50	0.89	0.9	0.99	0.81	0.99	0.27
ResNet18	0.88	0.9	0.98	0.81	0.98	0.31
Swin-s	0.86	0.88	0.99	0.77	**1**	0.04
ViT-b16	0.85	0.87	**1**	0.74	0.94	0.12
EfficientNet-B3	0.73	0.78	0.94	0.62	0.78	0.12

Having established fine-tuning as the preferred strategy, a broader set of experiments was performed, evaluating multiple classifier architectures from different families, including Convolutional Neural Networks like various ResNet and EfficientNet architectures, as well as Vision Transformer-based models such as ViT and Swin Transformer, spanning a range of model sizes. The comprehensive results are detailed in Table 4. Surprisingly, EfficientNet-B0 emerged as the top-performing model across all key evaluation metrics, despite being one of the smallest models tested. It achieved an impressive F1-score of 0.96, with high sensitivity for "easy" CPE cases (1.0) and a notable sensitivity of 0.77 for the challenging "hard" CPE cases. This performance significantly surpassed that of larger models like ResNet50, which, while achieving comparable sensitivity for "easy"

CPE, only managed a sensitivity of 0.27 for "hard" CPE, resulting in a lower overall F1-score of 0.89.

It is noteworthy that nearly all models demonstrated high specificity (close to 1.0), maintaining a low false-positive rate (i.e., uninfected samples being misclassified as CPE). This high specificity could be attributed to the distinct and relatively uniform visual characteristics of healthy, uninfected cell monolayers, which models may learn to differentiate effectively. It might also suggest that the features distinguishing infected from uninfected states are quite robust, even if detecting subtle infection (as in "hard" CPE) remains challenging.

Consistent with expectations for vision tasks on datasets of this scale, transformer-based models (ViT, Swin) generally did not match the performance levels of the top CNN architectures. This is likely because transformers, particularly when trained or fine-tuned on smaller, specialized datasets, often require more extensive data to effectively learn visual patterns compared to CNNs. CNNs possess inherent inductive biases (such as translation equivariance) beneficial for image analysis, which transformers typically lack, making the latter more data-hungry to achieve comparable performance levels.

The image classification experiments collectively demonstrate the general superiority of deep learning approaches over classical methods for CPE detection. However, it is also observed that the best classical model (Logistic Regression, F1 0.84) did manage to outperform some of the less effective deep learning configurations, such as EfficientNet-B3 (F1 0.73). This observation could indicate possible shortcut learning, since no augmentations were applied for classical methods compared to deep learning, or suboptimal deep learning performance could also be caused by a small dataset size and data complexity.

Table 5. Comparison of deep learning models for viral CPE detection with class weights.

				Sensitivity		
Model	F1	Accuracy	Specificity	All	Easy	Hard
EfficientNet_B0	**0.96**	**0.97**	**0.99**	**0.94**	**1**	**0.77**
EfficientNet-B0 class weights	0.89	0.9	0.92	0.89	0.96	0.65
ResNet50	**0.89**	**0.9**	**0.99**	0.81	**0.99**	0.27
ResNet50 class weights	0.88	0.89	0.94	**0.85**	0.98	**0.46**

After achieving superior performance with EfficientNet-B0, subsequent experiments aimed to further enhance its capabilities, particularly for the underrepresented "hard CPE" class, through class balancing. This involved assigning different weights to the classes (no CPE, easy CPE, hard CPE) within the loss function during training, as described in the Methods section 3.4. These experiments primarily focused on CNN models due to their demonstrated stability and strong performance on the dataset. However, the

outcomes of class balancing were inconclusive, as shown in Table 5. For instance, while applying class weights to ResNet50 led to an improvement in sensitivity for "hard" CPE cases, this came at the expense of performance on other classes or overall metrics, a common trade-off in imbalanced classification. Conversely, for the top-performing EfficientNet-B0, the introduction of class weights paradoxically led to a degradation in performance across nearly all categories.

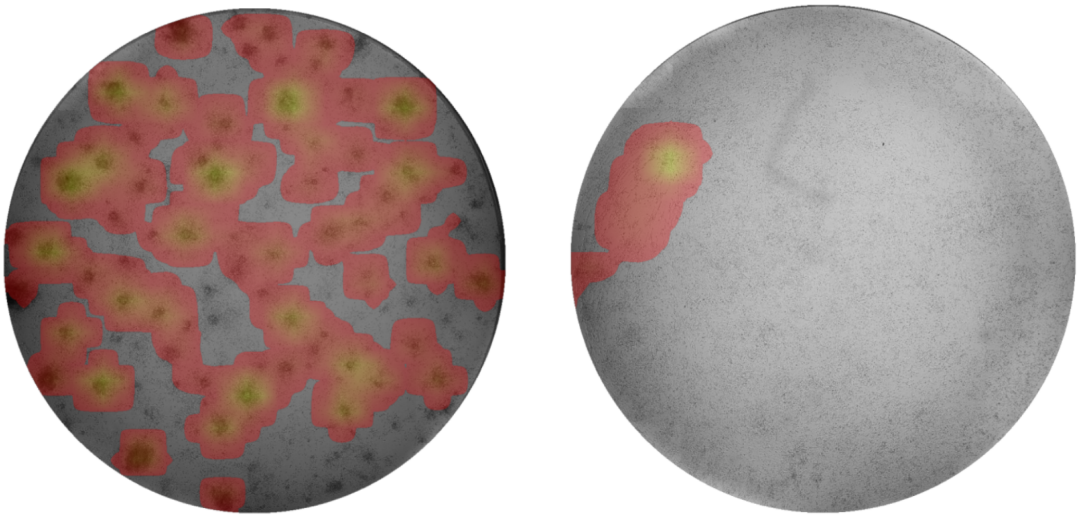


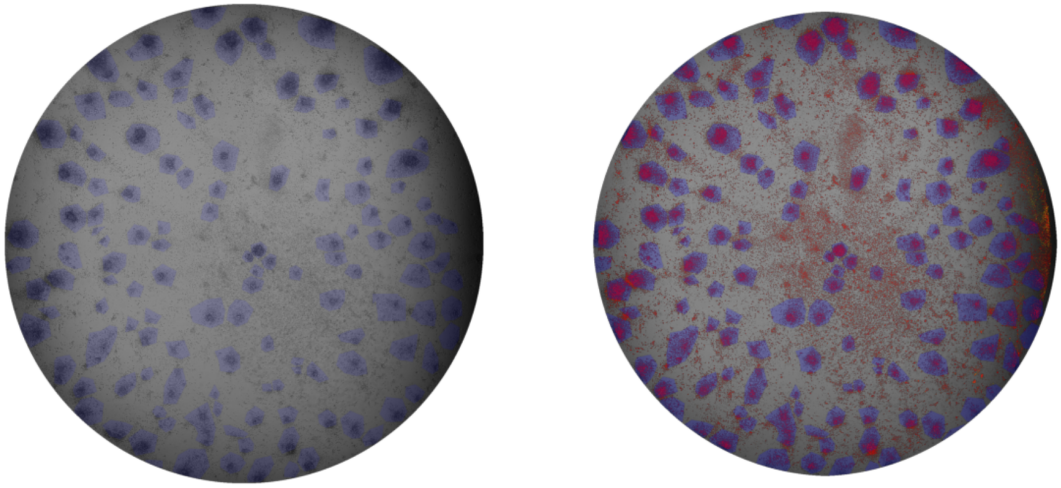
Figure 10. Grad-CAM++ values are colored in range from red to yellow for easy CPE (a) and hard CPE (b).

The exceptional performance metrics, particularly with EfficientNet-B0, raised the concern of potential shortcut learning. To investigate the decision-making process of the best-performing model, XAI techniques were employed. Various methods were applied to visualise the regions of an image that EfficientNet-B0 deemed most important for its predictions, using activations from its final convolutional block. Grad-CAM++ was found to produce particularly robust and interpretable saliency maps. Representative visualisations for test set images corresponding to CPE classifications are presented in Figure 10. These activation maps confirmed that the model was focusing on relevant cellular regions exhibiting morphological changes consistent with CPE, rather than relying on edge artifacts (as the image cropping was consistently applied). This successful application of explainable AI provided critical validation, increasing confidence that the model learned genuinely meaningful features and was not exploiting dataset biases or shortcuts. The results of Grad-CAM++ activations are also useful for CPE localisation, which will be explored further in the following section.

4.2 CPE Localization

Following image-level classification, the research progressed to the more granular task of CPE localisation, aiming to identify and localise specific regions within the images that exhibit cytopathic effects. This was pursued through both classical image processing techniques and various deep learning-based strategies.

4.2.1 Classical Segmentation



a) b)

Figure 11. Easy CPE case and local Otsu thresholding visualisation. Original image with CPE annotation in blue (a), combination of both with segmentation through Otsu in red (b).

Table 6. Comparative IoU scores of various segmentation approaches, including classical and deep learning-based methods, for localising Cytopathic Effect across all, "easy", and "hard" case categories.

	IoU		
Model	All	Easy	Hard
Local Otsu thresholding	0.04	0.06	0.03
HOG + Linear Model	0.04	0.05	0.02
EfficientNet-B0 GradCAM++	0.16	0.16	0.15
EfficientNet-B0 Unet	**0.57**	**0.60**	**0.34**

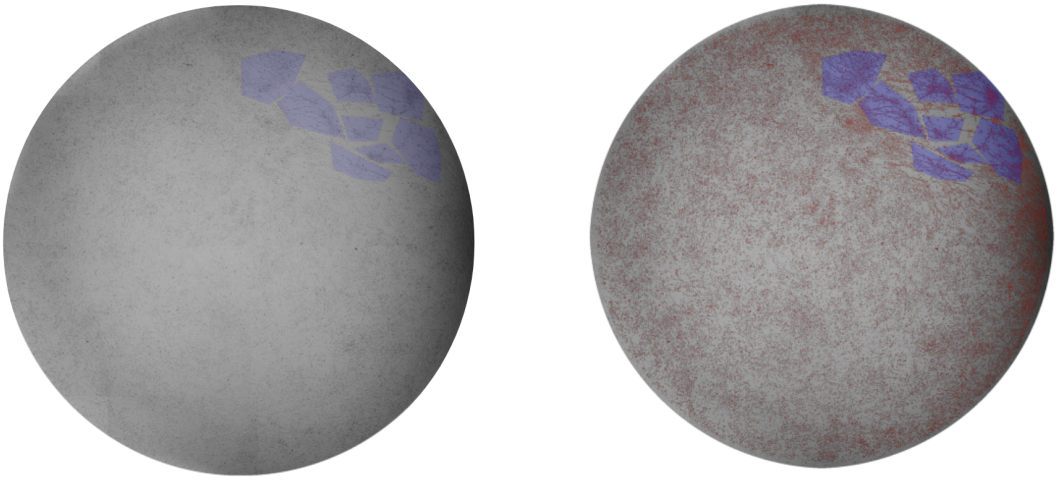
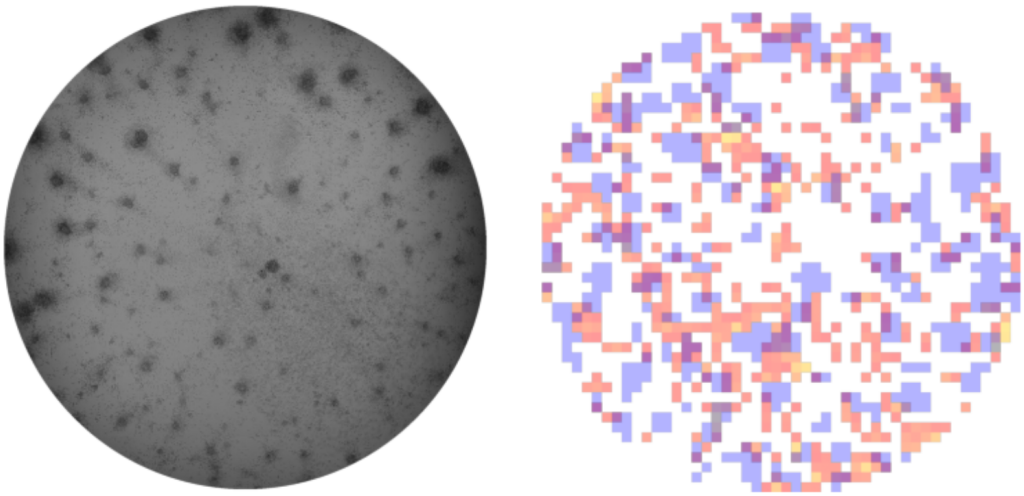


Figure 12. Hard CPE case and local Otsu thresholding visualisation. Original image with CPE annotation in blue (a), combination of both with segmentation through Otsu in red (b).

A range of classical image segmentation techniques was initially explored. One of the simplest, yet often effective for certain image types, is thresholding. Due to considerable brightness gradients observed across different images and even within different positions in a single well, global thresholding was unusable. Instead, local adaptive thresholding techniques were implemented to account for this variation. An initial approach combined local adaptive thresholding with Otsu's method for determining the optimal local threshold. However, as detailed in Table 6, the performance metrics for this method were uniformly poor across all CPE categories, with 0.04 IoU performance for all cases. Otsu-based local thresholding demonstrated some capability in segmenting "easy" CPE cases, which often manifest as large, dark cell aggregates that create a somewhat bimodal local histogram (Figure 11). This could potentially be improved via blob detection and an area filter. Conversely, for "hard" CPE cases, which are characterised by subtle changes in cell morphology rather than distinct intensity shifts, this method failed to identify CPE and instead tended to segment individual cells or background noise (Figure 12). This issue is due to the underlying assumption of thresholding methods, which rely on intensity differences that are not consistently present or distinct in "hard" CPE images. Another attempted method was the multiplication of standardised HOG features by the linear model trained on image classes. However, this also resulted in nearly random performance (see example for easy case on Figure 13.

Fully supervised classical segmentation approach was also attempted using HOG features. For this, ground truth annotation masks were downscaled to match the dimensionality of the extracted HOG features, and a model was trained to predict these mask



a) b)

Figure 13. Analysis of an "easy" CPE image: ground truth annotation and linear model feature importance derived from HOG predictions. (a) Original bright-field microscopy image of the 'Easy CPE' case. (b) Composite image displaying the ground truth CPE annotation (in blue) overlaid with a feature importance map. This map visualizes the model's assessment, derived from the sum of HOG features weighted by linear model coefficients, using a red (low summed value) to yellow (high summed value) gradient to highlight regions influential in the CPE classification. White areas (e.g., original background or white regions if pre-processed) indicate absence of CPE or low feature importance.

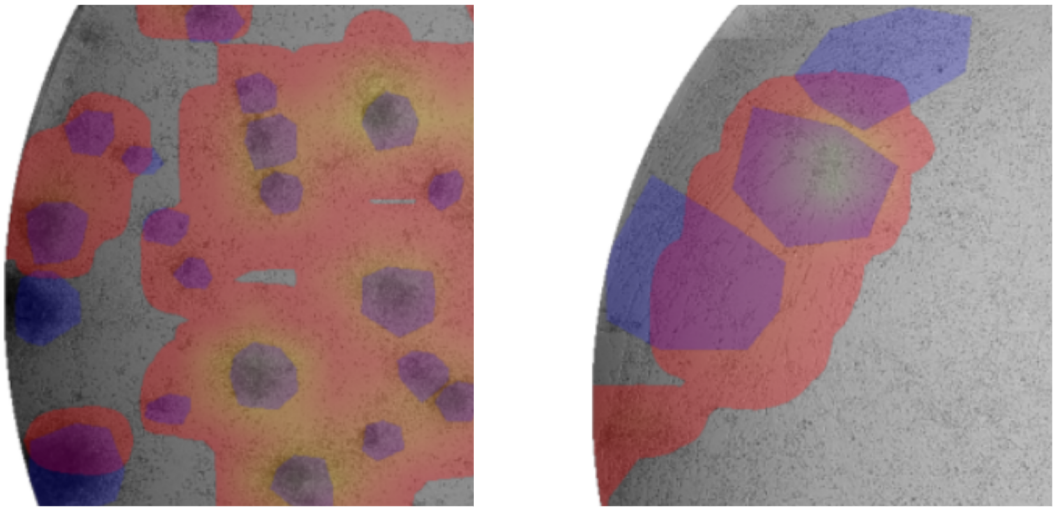
representations. While this HOG-based segmentation model was capable of overfitting to the training images, it failed to generalise to unseen images, yielding poor performance on the test set. Due to these limited initial results and pressing time constraints, further development of classical segmentation methods was discontinued in favour of exploring more promising deep learning approaches.

4.2.2 MIL

MIL approach encountered significant challenges: despite the application of data augmentation techniques, the model exhibited rapid overfitting to the training images. This is likely attributable to the inherent complexity of the transformer architecture relative to the limited size of our specialised microscopy dataset. Visualisations of the learned attention scores, which were intended to highlight important image regions, did not reveal a meaningful spatial correlation with actual CPE features. A much larger dataset is required for this approach to work. Given these issues of severe overfitting and

uninformative attention mechanisms in this context, this specific MIL approach was subsequently abandoned in favour of more directly supervised localisation methods.

4.2.3 Deep Learning Segmentation



a) b)

Figure 14. Grad-CAM++ visualization on top of the image with activation corresponding to CPE class. The ground truth annotation masks are highlighted in blue and Grad-CAM++ values are colored in red to yellow gradient. The lowest 10th percentile of Grad-CAM++ values is masked. a) showcases the easy CPE case, b) is the hard CPE case.

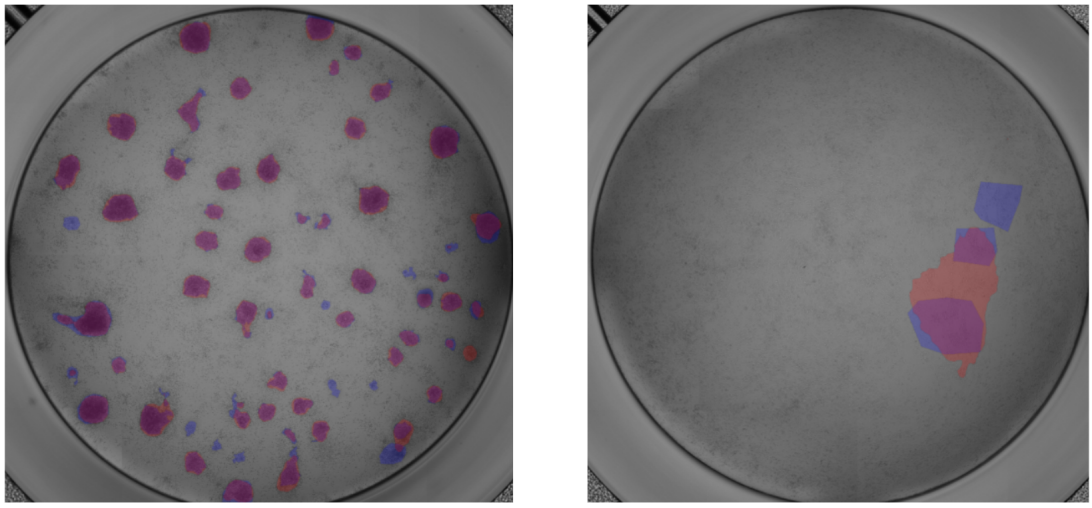
The initial deep learning segmentation of images was performed in an unsupervised manner. This was based on the Grad-CAM+ thresholding. The threshold value was chosen to maximize IoU.

An initial exploration into deep learning-based segmentation was performed by classification model explainability thresholding (i.e., without using segmentation masks for training this specific step). This method used the class activation maps generated by the previously trained high-performing image classifier (EfficientNet-B0). Specifically, Grad-CAM++ saliency maps, which highlight image regions most influential for the classifier's CPE prediction, were generated. These maps were then thresholded to produce binary segmentation masks, with the threshold value optimised to maximise IoU against a validation subset of ground truth masks.

Visual examples of this technique on test set images are presented in Figure 14a for an "easy" CPE case. The resulting saliency map from Grad-CAM++ often showed a

notable visual resemblance to the actual ground truth annotations that were subsequently used for training supervised segmentation models. More strikingly, as shown in Figure 14b, even for some "hard" CPE cases, the activation maps appeared to correctly localise regions of subtle infection. This was a somewhat surprising outcome, given the limited number of 'hard CPE' examples in the original classification training set, suggesting the classifier had learned some genuinely subtle discriminative features. Despite these encouraging qualitative observations, the quantitative performance of this thresholded Grad-CAM++ approach was modest, as detailed in Table 6. The overall IoU plateaued around 0.16, which is relatively low. This was largely attributed to the nature of CAMs, which tend to highlight broader regions of interest rather than producing tightly bounded segmentations; the activations often extended well beyond the precise ground truth boundaries, thereby "diluting" the union term in the IoU calculation and reducing the score.

Thresholding and metrics calculation based on the Grad-CAM++ is referenced in Table 6. The IoU metric for all cases is not great, around 0.16, and is largely limited because activations go far beyond ground truth labels, which dilutes the union.



a) b)

Figure 15. Segmentation results from the U-Net model with EfficientNet-B0 backbone on test images, comparing model predictions against ground truth annotations for CPE. Ground truth masks are depicted in blue, model-predicted masks in red, and areas where these overlap are in purple. (a) Visualisation for an easy CPE case. (b) Visualisation for a hard CPE case.

Following the insights from unsupervised methods, fully supervised deep learning segmentation was performed using the well-established U-Net architecture. To leverage learned features and potentially accelerate training, the U-Net's encoder was initialized

with pre-trained weights from the EfficientNet-B0 model on ImageNet, which had demonstrated the best performance in the classification task. Representative segmentation outputs from this model on test set images are shown in Figure 15. Quantitatively, the U-Net model achieved a mean IoU of approximately 0.60 for "easy" CPE cases, a substantial improvement over classical and unsupervised deep learning approaches. However, performance on "hard" CPE cases was considerably lower, with a mean IoU of around 0.34.

A significant factor potentially influencing these IoU scores, particularly for "hard" CPE, is the challenge and subjectivity in annotating subtle biological phenomena. To illustrate this, an internal consistency check was performed where the primary annotator (author) re-labelled the same image after a period of several weeks, resulting in an IoU of approximately 0.7 between the two sets of self-generated annotations for visible CPE structures. This highlights the intrinsic difficulty in achieving perfect pixel-wise annotation consistency, especially without multiple expert, cross-verified annotations. These data quality considerations, originating from the subjective visual assessment by a single annotator for the initial segmentation masks, likely impose a practical upper limit on achievable segmentation performance and represent a key limitation of the current dataset.

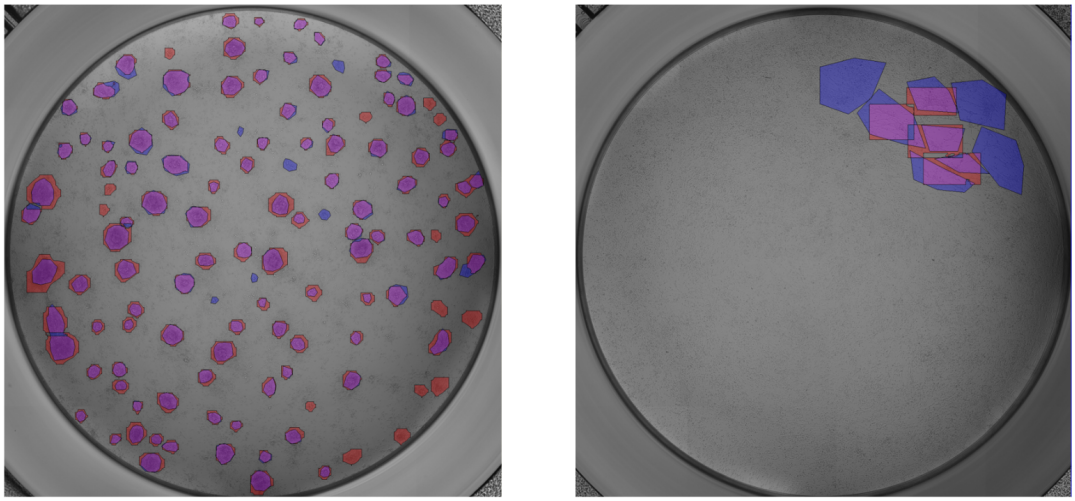
To further refine the localization and enable the delineation of individual CPE instances, instance segmentation was approached using YOLO-family models. It is important to note that the dataset for instance segmentation was iteratively improved; initial annotations were supplemented over time as more predictions were made, manually reviewed, and corrected. Early experiments with larger, general-purpose pre-trained YOLO models exhibited significant overfitting to our specific CPE dataset and did not generalize well. To rigorously evaluate models under these conditions and select the best performers, a fixed, representative test set was established.

Table 7. Results of segmentation model.

	Instance Precision			Instance Sensitivity			IoU		
Model Type	All	Easy	Hard	All	Easy	Hard	All	Easy	Hard
Yolo11s	0.82	0.83	**0.62**	**0.61**	**0.63**	**0.17**	**0.45**	**0.49**	**0.15**
Yolo8m	**0.83**	**0.84**	0.17	0.22	0.23	0.02	0.43	0.47	0.09

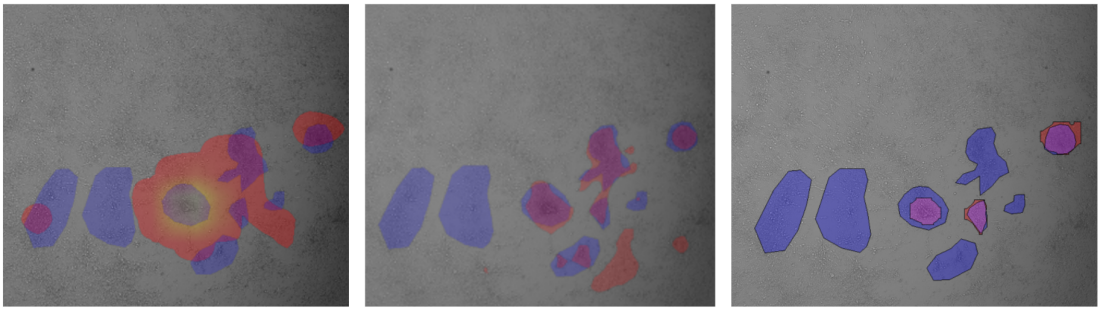
Among the evaluated configurations, models denoted as YOLOv11s and YOLOv8m were shortlisted as the most promising. Notably, the smaller YOLOv11s model ultimately outperformed the larger YOLOv8m variant, as detailed in Table 7. This outcome may suggest that for this specialized dataset with subtle and sometimes limited features, a smaller, more agile model architecture was less prone to overfitting and better able to capture the relevant characteristics of CPE instances. The sample from the test set is visualized in Figure 16.

To visually compare the efficacy of the investigated deep learning approaches on challenging instances, Figure 17 presents the respective segmentation outputs all applied to a representative hard CPE case selected from the test set.



a) b)

Figure 16. Instance segmentation results from the YOLOv11s model on representative test images, comparing model predictions against ground truth annotations for CPE. Ground truth masks are depicted in blue, model-predicted masks in red, and areas where these overlap are in purple. (a) Visualisation for an easy CPE case. (b) Visualisation for a hard CPE case.



a) b) c)

Figure 17. Segmentation of a challenging hard CPE case using various approaches. (a) Unsupervised segmentation generated by thresholding Grad-CAM++ activations from an EfficientNet-B0 classifier. (b) Supervised semantic segmentation produced by a U-Net architecture with an EfficientNet-B0 encoder. (c) Instance segmentation predictions yielded by a YOLOv11s model.

5 Conclusion

An important aspect in virological research and diagnostics is the observation of CPE — the spectrum of morphological alterations induced in cells by viral replication. This serves as a visual indicator of infection in cell cultures. While the unique characteristics of CPE can be indicative of specific virus-cell interactions, the traditional reliance on manual brightfield microscopy inspection for its detection is inherently labor-intensive, time-consuming, and susceptible to inter-observer subjectivity, presenting a considerable bottleneck in an era demanding rapid and high-throughput analytical capabilities.

This thesis addressed these critical limitations by developing, implementing, and evaluating a set of computer vision methodologies for the automated detection and, crucially, the detailed quantification of CPE induced by xenotrophic murine leukemia virus, which is a model retrovirus. Our investigation systematically progressed from foundational classical image analysis and machine learning techniques to advanced deep learning architectures. High-performing classifiers, such as EfficientNet-B0, were identified for image CPE status determination, with their decision-making processes validated through Explainable AI techniques like Grad-CAM++, which also helped confirm the successful mitigation of shortcut learning. While Multiple Instance Learning showed limitations due to dataset size and model complexity, the core contribution of this work lies in its in-depth exploration and application of instance segmentation techniques for granular CPE characterization and quantification. This focus on moving beyond binary classification to provide quantitative, spatially-resolved data on CPE extent represents a significant extension of prior automated efforts in the field, which have predominantly centered on classification.

Key findings demonstrated the superiority of deep learning approaches for both classification and segmentation tasks, particularly in distinguishing subtle, "hard-to-detect" CPE manifestations from uninfected cells or "easy" CPE cases. The U-Net architecture and YOLO-family models successfully segmented CPE, with smaller, optimized YOLO variants proving effective for instance-level predictions. This research also highlighted the impact of annotation quality and consistency on model performance, underscoring the challenges of ground truth generation for subtle biological phenomena, as evidenced by intra-annotator variability assessments.

A practical strategy emerging from this work involves a hierarchical approach: leveraging high-performance classifiers to rapidly identify infected samples, followed by targeted application of segmentation models to these positive cases for detailed quantification, thereby optimizing computational resources.

Future research could build upon these foundations. Expanding the training datasets with more diverse images, potentially encompassing different cell lines, viruses, or imaging conditions, and enriched with multi-expert, cross-verified annotation masks, would enhance model robustness and generalizability. With such augmented datasets, revisiting approaches like Multiple Instance Learning could prove useful. Further steps

include exploring advanced self-supervised or few-shot learning paradigms to reduce annotation dependency, and prospectively validating these automated systems in real-world research settings.

In conclusion, this thesis demonstrates the potential of computer vision, particularly deep learning-driven applications, to automate and refine the analysis of viral CPE. By delivering tools for both detection and quantification of x-MuLV-induced effects, this work contributes valuable methodologies and insights towards enhancing the efficiency and depth of virological investigations, aiding in future efforts to understand and combat viral diseases.

6 Aknowledgements

I am sincerely grateful to my supervisor, Dmytro Fishman, for his unparalleled guidance, inspiration and understanding. General compliment to all of the colleagues from Biomedical Computer Vision lab of the University of Tartu who helped numerous times by guiding to answers through technical questions and sanity checks.

References

[1] Meng Lv et al. “Coronavirus disease (COVID-19): a scoping review”. In: *Eurosurveillance* 25.15 (Apr. 2020). ISSN: 1560-7917. DOI: 10.2807/1560-7917.es.2020.25.15.2000125. URL: http://dx.doi.org/10.2807/1560-7917.ES.2020.25.15.2000125.

[2] Swetha Vijayakrishnan, Yaming Jiu, and J Robin Harris. *Virus infected cells*. en. Cham, Switzerland: Springer Nature, Dec. 2023.

[3] Daniel Céspedes-Tenorio and Jorge L. Arias-Arias. “The Virus-Induced Cytopathic Effect”. In: *Virus Infected Cells*. Springer International Publishing, 2023, pp. 197–210. ISBN: 9783031400865. DOI: 10.1007/978-3-031-40086-5_7. URL: http://dx.doi.org/10.1007/978-3-031-40086-5_7.

[4] Anthony Petkidis et al. “A versatile automated pipeline for quantifying virus infectivity by label-free light microscopy and artificial intelligence”. In: *Nature Communications* 15.1 (June 2024). ISSN: 2041-1723. DOI: 10.1038/s41467-024-49444-1. URL: http://dx.doi.org/10.1038/s41467-024-49444-1.

[5] Michael A. Arbib. *The Handbook of Brain Theory and Neural Networks*. 2nd. Cambridge, MA, USA: MIT Press, 2002. ISBN: 0262011972.

[6] Michael Gadermayr and Maximilian Tschuchnig. “Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations amp; future potential”. In: *Computerized Medical Imaging and Graphics* 112 (Mar. 2024), p. 102337. ISSN: 0895-6111. DOI: 10.1016/j.compmedimag.2024.102337. URL: http://dx.doi.org/10.1016/j.compmedimag.2024.102337.

[7] Alan Rein. “Murine Leukemia Viruses: Objects and Organisms”. In: *Advances in Virology* 2011 (2011), pp. 1–14. ISSN: 1687-8647. DOI: 10.1155/2011/403419. URL: http://dx.doi.org/10.1155/2011/403419.

[8] Zeynep Akkutay-Yoldar et al. “A web-based artificial intelligence system for label-free virus classification and detection of cytopathic effects”. In: *Scientific Reports* 15.1 (Feb. 2025). ISSN: 2045-2322. DOI: 10.1038/s41598-025-89639-0. URL: http://dx.doi.org/10.1038/s41598-025-89639-0.

[9] Rupert Dodkins et al. “A rapid, high-throughput, viral infectivity assay using automated brightfield microscopy with machine learning”. In: *SLAS Technology* 28.5 (Oct. 2023), pp. 324–333. ISSN: 2472-6303. DOI: 10.1016/j.slast.2023.07.003. URL: https://doi.org/10.1016/j.slast.2023.07.003.

[10] Ting-En Wang et al. “Differentiation of Cytopathic Effects (CPE) induced by influenza virus infection using deep Convolutional Neural Networks (CNN)”. In: *PLOS Computational Biology* 16.5 (May 2020). Ed. by Amber M. Smith, e1007883. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1007883. URL: http://dx.doi.org/10.1371/journal.pcbi.1007883.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[12] Ross B. Girshick et al. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In: *CoRR* abs/1311.2524 (2013). arXiv: 1311.2524. URL: http://arxiv.org/abs/1311.2524.

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *CoRR* abs/1505.04597 (2015). arXiv: 1505.04597. URL: http://arxiv.org/abs/1505.04597.

[14] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *nature* 521.7553 (2015), pp. 436–444.

[15] Karla Cristine C. DOYSABAS et al. “ATeam technology for detecting early signs of viral cytopathic effect”. In: *Journal of Veterinary Medical Science* 82.3 (2020), pp. 387–393. ISSN: 1347-7439. DOI: 10.1292/jvms.20-0021. URL: http://dx.doi.org/10.1292/jvms.20-0021.

[16] Efstathia Papafragkou et al. “Challenges of Culturing Human Norovirus in Three-Dimensional Organoid Intestinal Cell Culture Models”. In: *PLoS ONE* 8.6 (June 2013). Ed. by Amit Kapoor, e63485. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0063485. URL: http://ddoi.org/10.1371/journal.pone.0063485.

[17] Alan Baer and Kylene Kehn-Hall. “Viral Concentration Determination Through Plaque Assays: Using Traditional and Novel Overlay Systems”. In: *Journal of Visualized Experiments* 93 (Nov. 2014). ISSN: 1940-087X. DOI: 10.3791/52065. URL: http://dx.doi.org/10.3791/52065.

[18] Neha Goswami et al. “Label-free SARS-CoV-2 detection and classification using phase imaging with computational specificity”. In: *Light: Science amp; Applications* 10.1 (Sept. 2021). ISSN: 2047-7538. DOI: 10.1038/s41377-021-00620-8. URL: http://dx.doi.org/10.1038/s41377-021-00620-8.

[19] Zili Li, Margaret Blair, and Lauren Thorner. “PG-4 cell plaque assay for xenotropic murine leukemia virus”. In: *Journal of Virological Methods* 81.1–2 (Aug. 1999), pp. 47–53. ISSN: 0166-0934. DOI: 10.1016/s0166-0934(99)00064-6. URL: http://dx.doi.org/10.1016/s0166-0934(99)00064-6.

[20] Niall O' Mahony et al. "Deep Learning vs. Traditional Computer Vision". In: *CoRR* abs/1910.13796 (2019). arXiv: 1910.13796. URL: http://arxiv.org/abs/1910.13796.

[21] Nobuyuki Otsu. "A Threshold Selection Method from Gray-Level Histograms". In: *IEEE Transactions on Systems, Man, and Cybernetics* 9.1 (1979), pp. 62–66. DOI: 10.1109/TSMC.1979.4310076.

[22] John Canny. "A computational approach to edge detection". In: *IEEE Transactions on pattern analysis and machine intelligence* 6 (1986), pp. 679–698.

[23] Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. "Design of an image edge detection filter using the Sobel operator". In: *IEEE Journal of solid-state circuits* 23.2 (1988), pp. 358–367.

[24] N. Dalal and B. Triggs. "Histograms of oriented gradients for human detection". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. 2005, 886–893 vol. 1. DOI: 10.1109/CVPR.2005.177.

[25] Junyi Chai et al. "Deep learning in computer vision: A critical review of emerging techniques and application scenarios". In: *Machine Learning with Applications* 6 (Dec. 2021), p. 100134. ISSN: 2666-8270. DOI: 10.1016/j.mlwa.2021.100134. URL: http://dx.doi.org/10.1016/j.mlwa.2021.100134.

[26] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[27] Tsung-Yi Lin et al. "Microsoft COCO: Common Objects in Context". In: *CoRR* abs/1405.0312 (2014). arXiv: 1405.0312. URL: http://arxiv.org/abs/1405.0312.

[28] Andre Esteva et al. "Deep learning-enabled medical computer vision". In: *npj Digital Medicine* 4.1 (Jan. 2021), p. 5. ISSN: 2398-6352. DOI: 10.1038/s41746-020-00376-2. URL: https://doi.org/10.1038/s41746-020-00376-2.

[29] Warren S McCulloch and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5.4 (1943), pp. 115–133.

[30] Laveen N. Kanal. "Perceptron". In: *Encyclopedia of Computer Science*. GBR: John Wiley and Sons Ltd., 2003, pp. 1383–1385. ISBN: 0470864125.

[31] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), pp. 533–536.

[32] Y LeCun et al. “Proc. advances in neural information processing systems”. In: *Proc. advances in neural information processing systems* (1990).

[33] Isaac Westby et al. “A Design on Multilayer Perceptron (MLP) Neural Network for Digit Recognition”. In: *Advances in Artificial Intelligence and Applied Cognitive Computing*. Ed. by Hamid R. Arabnia et al. Cham: Springer International Publishing, 2021, pp. 729–741. ISBN: 978-3-030-70296-0.

[34] Preetum Nakkiran et al. “Deep Double Descent: Where Bigger Models and More Data Hurt”. In: *CoRR* abs/1912.02292 (2019). arXiv: 1912.02292. URL: http://arxiv.org/abs/1912.02292.

[35] Liqun Hou and Zijing Li. “Fault Diagnosis of Rolling Bearing Based on Tunable Q-Factor Wavelet Transform and Convolutional Neural Network”. In: *International Journal of Online and Biomedical Engineering (iJOE)* 16.02 (Feb. 2020), pp. 47–61. ISSN: 2626-8493. DOI: 10.3991/ijoe.v16i02.11953. URL: http://dx.doi.org/10.3991/ijoe.v16i02.11953.

[36] David H Hubel and Torsten N Wiesel. “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex”. In: *The Journal of physiology* 160.1 (1962), p. 106.

[37] Jason Sanders and Edward Kandrot. *CUDA by Example: An Introduction to General-Purpose GPU Programming*. 1st. Addison-Wesley Professional, 2010. ISBN: 0131387685.

[38] Kaiming He et al. “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”. In: *CoRR* abs/1502.01852 (2015). arXiv: 1502.01852. URL: http://arxiv.org/abs/1502.01852.

[39] Andre Esteva et al. “Dermatologist-level classification of skin cancer with deep neural networks”. In: *nature* 542.7639 (2017), pp. 115–118.

[40] Razvan Pascanu, Tomás Mikolov, and Yoshua Bengio. “Understanding the exploding gradient problem”. In: *CoRR* abs/1211.5063 (2012). arXiv: 1211.5063. URL: http://arxiv.org/abs/1211.5063.

[41] Christian Szegedy et al. “Going Deeper with Convolutions”. In: *CoRR* abs/1409.4842 (2014). arXiv: 1409.4842. URL: http://arxiv.org/abs/1409.4842.

[42] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

[43] Mingxing Tan and Quoc V. Le. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. 2020. arXiv: 1905.11946 [cs.LG]. URL: https://arxiv.org/abs/1905.11946.

[44] Mingxing Tan and Quoc V. Le. *EfficientNetV2: Smaller Models and Faster Training*. 2021. arXiv: 2104.00298 [cs.CV]. URL: https://arxiv.org/abs/2104.00298.

[45] Fabian Isensee et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature Methods* 18.2 (Dec. 2020), pp. 203–211. ISSN: 1548-7105. DOI: 10.1038/s41592-020-01008-z. URL: http://dx.doi.org/10.1038/s41592-020-01008-z.

[46] Ange Lou, Shuyue Guan, and Murray Loew. *DC-UNet: Rethinking the U-Net Architecture with Dual Channel Efficient CNN for Medical Images Segmentation*. 2020. arXiv: 2006.00414 [eess.IV]. URL: https://arxiv.org/abs/2006.00414.

[47] Nabil Ibtehaz and M. Sohel Rahman. "MultiResUNet : Rethinking the U-Net Architecture for Multimodal Biomedical Image Segmentation". In: *CoRR* abs/1902.04049 (2019). arXiv: 1902.04049. URL: http://arxiv.org/abs/1902.04049.

[48] Kaiming He et al. "Mask R-CNN". In: *CoRR* abs/1703.06870 (2017). arXiv: 1703.06870. URL: http://arxiv.org/abs/1703.06870.

[49] Joseph Redmon et al. "You Only Look Once: Unified, Real-Time Object Detection". In: *CoRR* abs/1506.02640 (2015). arXiv: 1506.02640. URL: http://arxiv.org/abs/1506.02640.

[50] Ashish Vaswani et al. "Attention Is All You Need". In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[51] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *CoRR* abs/2010.11929 (2020). arXiv: 2010.11929. URL: https://arxiv.org/abs/2010.11929.

[52] Hugo Touvron et al. *Training data-efficient image transformers distillation through attention*. 2021. arXiv: 2012.12877 [cs.CV]. URL: https://arxiv.org/abs/2012.12877.

[53] Ze Liu et al. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. 2021. arXiv: 2103.14030 [cs.CV]. URL: https://arxiv.org/abs/2103.14030.

[54] Byeongho Heo et al. *Rethinking Spatial Dimensions of Vision Transformers*. 2021. arXiv: 2103.16302 [cs.CV]. URL: https://arxiv.org/abs/2103.16302.

[55] Fuzhen Zhuang et al. *A Comprehensive Survey on Transfer Learning*. 2020. arXiv: 1911.02685 [cs.LG]. URL: https://arxiv.org/abs/1911.02685.

[56] Benedikt Roth et al. “Low-Resource Finetuning of Foundation Models Beats State-of-the-Art in Histopathology”. In: *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, May 2024, pp. 1–5. DOI: 10.1109/isbi56570.2024.10635695. URL: http://dx.doi.org/10.1109/ISBI56570.2024.10635695.

[57] David Picard. “Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision”. In: *CoRR* abs/2109.08203 (2021). arXiv: 2109.08203. URL: https://arxiv.org/abs/2109.08203.

[58] Zimeng Lyu et al. *An Experimental Study of Weight Initialization and Weight Inheritance Effects on Neuroevolution*. 2020. arXiv: 2009.09644 [cs.NE]. URL: https://arxiv.org/abs/2009.09644.

[59] Ramprasaath R. Selvaraju et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74.

[60] Aditya Chattopadhyay et al. “Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks”. In: *CoRR* abs/1710.11063 (2017). arXiv: 1710.11063. URL: http://arxiv.org/abs/1710.11063.

[61] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. “"Why Should I Trust You?": Explaining the Predictions of Any Classifier”. In: *CoRR* abs/1602.04938 (2016). arXiv: 1602.04938. URL: http://arxiv.org/abs/1602.04938.

[62] Scott M. Lundberg and Su-In Lee. “A unified approach to interpreting model predictions”. In: *CoRR* abs/1705.07874 (2017). arXiv: 1705.07874. URL: http://arxiv.org/abs/1705.07874.

[63] Mathilde Caron et al. “Emerging Properties in Self-Supervised Vision Transformers”. In: *CoRR* abs/2104.14294 (2021). arXiv: 2104.14294. URL: https://arxiv.org/abs/2104.14294.

[64] Maxime Oquab et al. *DINOv2: Learning Robust Visual Features without Supervision*. 2024. arXiv: 2304.07193 [cs.CV]. URL: https://arxiv.org/abs/2304.07193.

[65] Feng Li et al. *Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and Segmentation*. 2022. arXiv: 2206.02777 [cs.CV]. URL: https://arxiv.org/abs/2206.02777.

[66] Zhixiong Nan et al. *DI-MaskDINO: A Joint Object Detection and Instance Segmentation Model*. 2024. arXiv: 2410.16707 [cs.CV]. URL: https://arxiv.org/abs/2410.16707.

[67] Maxim Tkachenko et al. *Label Studio: Data labeling software*. Open source software available from https://github.com/heartexlabs/label-studio. 2020-2022. URL: https://github.com/heartexlabs/label-studio.

[68] Alexander Kirillov et al. *Segment Anything*. 2023. arXiv: 2304.02643 [cs.CV]. URL: https://arxiv.org/abs/2304.02643.

[69] Stefan Van der Walt et al. "scikit-image: image processing in Python". In: *PeerJ* 2 (2014), e453.

[70] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. *Attention-based Deep Multiple Instance Learning*. 2018. arXiv: 1802.04712 [cs.LG]. URL: https://arxiv.org/abs/1802.04712.

Appendix

I. Glossary

Histogram of Oriented Gradients (HOG): A feature descriptor that is a computer vision technique used to detect objects by counting occurrences of gradient orientation in localized portions of an image. It's particularly effective for capturing edge or shape information.
96-well plate: A flat plate with 96 small, circular depressions (**wells**), commonly used in biological assays for culturing cells and conducting multiple experiments simultaneously, such as viral infection studies with varying conditions or concentrations.
Bright-field Microscopy: A standard optical microscopy technique where light is transmitted through a specimen, and contrast in the image comes from the differential absorption or scattering of light by parts of the specimen.
CAM (Class Activation Mapping): A visualization technique used in deep learning to identify which regions in an input image are most influential for a CNN's classification decision for a specific class.
Cell Culture: The process of growing cells under controlled conditions, typically outside their natural environment, often in laboratory vessels like 96-well plates. Used to study cellular processes, including viral infections.
Cell Line: A population of cells derived from a single cell and grown in culture, capable of stable propagation over many generations.
CPE (Cytopathic Effect): Morphological changes in host cells resulting from viral infection, observable under a microscope. These changes can include cell rounding, shrinkage, detachment, syncytia formation, and the appearance of inclusion bodies, serving as a key visual indicator of viral presence and activity.
HOG (Histogram of Oriented Gradients): A feature descriptor used in computer vision for object detection and image classification, which counts occurrences of gradient orientation in localized portions of an image.
Saliency Map: A visualization that highlights the regions in an input image that are most important or influential for a model's output or decision (e.g., CAMs).
Shortcut Learning: A phenomenon where a machine learning model learns to make predictions based on unintended or spurious correlations (shortcuts) in the training data, rather than learning the true underlying patterns, often leading to poor generalization.
Viral Load: A measure of the quantity of virus in a given volume, often correlated with the extent or severity of CPE.
Viral Titre: A measure of the concentration of infectious virus particles in a sample.
Well (in a 96-well plate): One of the small, circular depressions in a multi-well plate used for individual cell cultures or reactions.

II. Supplementary figures

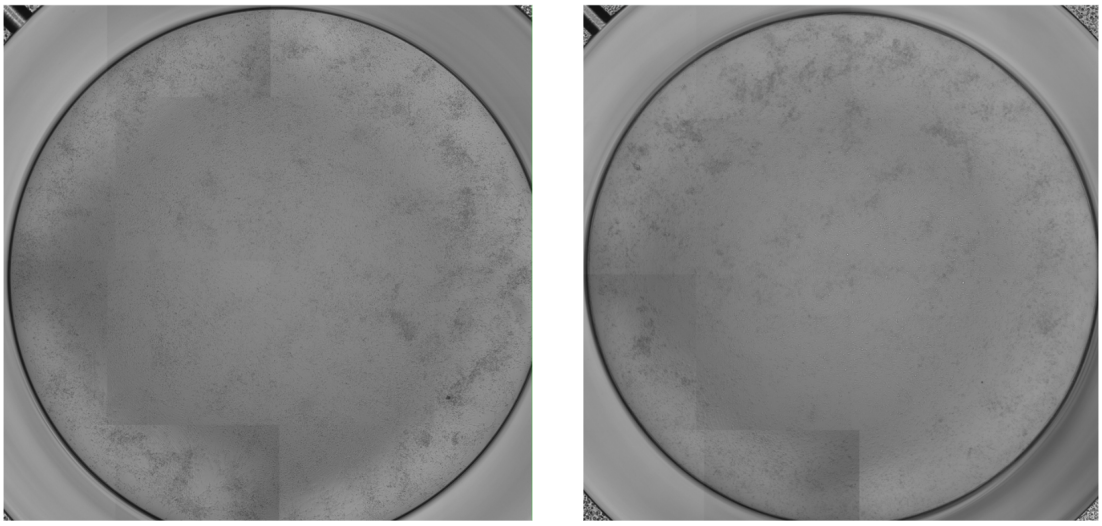


Figure .1. Examples of images from excluded plates. Large heterogeneity of image stitching is present.

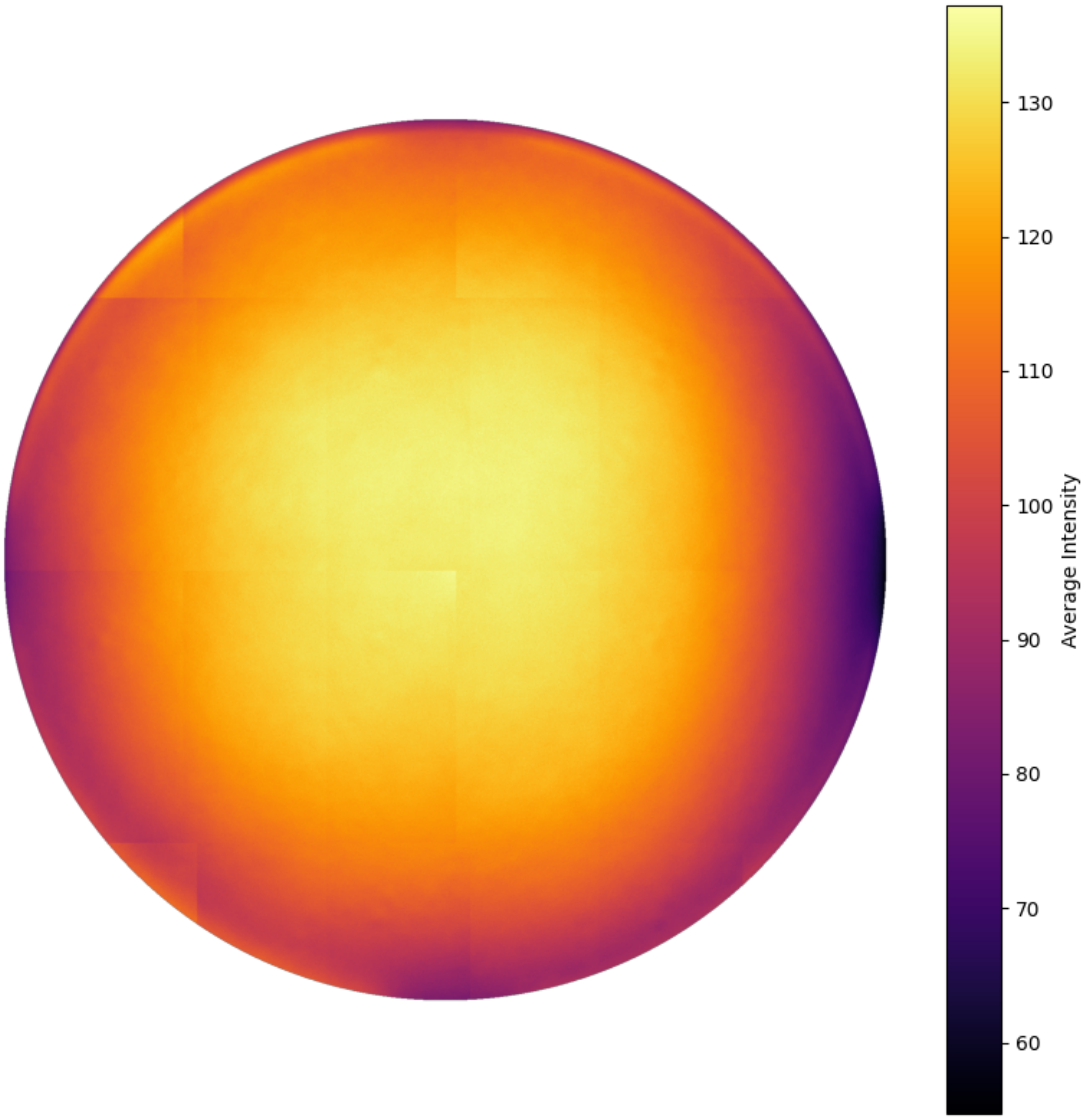


Figure .2. The average pixel intensity of images selected for the analysis. The brightness gradient is different for different columns of the plate.

III. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Aleksandr Makarov**,
*(*author's name)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Mastering the Unseen: Approaches to Hard-to-Detect Viral Cytopathic Effect,
   *(*title of thesis)

   supervised by Dmytro Fishman.
   *(*supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Aleksandr Malarov
15/05/2025