UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Technology

Ali Zeynalli

Computer Vision Meets Microbiology: Deep Learning Algorithms for Classifying Cell Treatments in Microscopy Images

Bachelor's Thesis (12 ECTS)

Curriculum Science and Technology

Supervisor:
Assistant Professor, Dmytro Fishman (PhD)

Tartu 2023

Computer Vision Meets Microbiology: Deep Learning Algorithms for Classifying Cell Treatments in Microscopy Images

Abstract:

Cell classification is one of the most complex challenges in cellular research that has significant importance to personalised medicine, cancer diagnostics and disease prevention. The accurate classification of cells based on their unique characteristics provides valuable insights into a patient's health status and in guiding treatment decisions. Thanks to recent technological advancements, cellular research has experienced significant progress in the use of deep learning and has become a valuable tool for tackling complicated tasks such as cell classification. In this study, we explored the capability of state-of-the-art deep learning models such as ResNet, ViT and Swin Transformer to automatically classify brightfield and fluorescent microscopy images across single and multiple channels into four cell treatments: Palbociclib, MLN8237, AZD1152, and CYC116. The results have revealed that Swin Transformer surpasses the other models for cell treatment classification on multi-channel fluorescent and brightfield images, achieving the highest accuracy of 86% and 59%, correspondingly. However, the highest accuracy achieved on single-channel brightfield images was 61%, using the ResNet-50 model. The previous research has shown that combining multiple channels yields better performance which necessitates further investigation into the capacity of deep learning models for automating the cell treatment classification of single- and multi-channel brightfield microscopy images.

Keywords:

Machine learning, Deep learning, Neural networks, Image classification

CERCS:

B110 Bionformatics, medical informatics,Biomathematics,Biometrics

P176 Artitifical intelligence

T111 Imaging, Image processing

Arvutinägemine kohtub mikrobioloogiaga: süvaõppe algoritmid rakuravi klassifitseerimiseks mikroskoopiapiltidel

Lühikokkuvõte:

Rakkude klassifitseerimine on üks rakuuuringute keerukamaid väljakutseid, millel on eriti suur tähtsus isikupärastatud meditsiinis, vähi diagnostikas ja haiguste ennetamisel. Rakkude täpne liigitamine nende ainulaadsete tunnuste põhjal annab väärtuslikku teavet patsiendi tervisliku seisundi kohta ja aitab langetada raviotsuseid. Tänu hiljutistele tehnoloogilistele edusammudele on rakuuuringud süvaõppe kasutamisel saavutanud märkimisväärset edu ja muutunud väärtuslikuks vahendiks keeruliste toimingute, näiteks rakkude klassifitseerimise, teostamisel. Selles uuringus uurisime tipptasemel süvaõppe mudelite (nt ResNet, ViT ja Swin Transformer) võimet klassifitseerida helevälja ja fluorestsentsmikroskoopia kujutisi ühe ja mitme kanaliga automaatselt neljaks rakutöötluseks: Palbociclib, MLN8237, AZD1152. ja CYC116. Tulemused on näidanud, et Swin Transformer ületab mitme kanaliga fluorestsents- ja helevälja kujutiste puhul rakutöötluse klassifitseerimise teisi mudeleid, saavutades kõrgeima täpsuse, vastavalt 86% ja 59%. ResNet-50 mudelit kasutades oli ühe kanaliga erevälja kujutisel saavutatud kõrgeim täpsus aga 61%. Varasemad uuringud on näidanud, et mitme kanali kombineerimine annab parema soorituse. Järelikult nõuab süvaõppemudelite suutlikkus täiendavat uurimist, et automatiseerida ühe- ja mitmekanaliliste helevälja mikroskoopiakujutiste rakutöötluse klassifitseerimist.

Võtmesõnad:

Masinõpe, Süvaõpe, Närvivõrgud, Piltide klassifitseerimine

CERCS:

B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

P176 Tehisintellekt

T111 Pilditehnika

TABLE OF CONTENTS

TERMS, ABBREVIATIONS AND NOTATIONS

AI - Artificical Intelligence

CNN - Convolutional Neural Network

ML - Machine Learning

DL - Deep Learning

RGB - Red, Green, Blue

ResNet - Residual Network

ViT - Vision Transformer

INTRODUCTION

All living things are composed of cells, which are the smallest unit of life. Understanding the structural organization of cells is a requirement for comprehending the function of cells [1]. Acknowledging how cells function explains biological characteristics that maintain our life while also revealing novel treatments for disease [2]. Despite the fact that cellular research has already contributed to cancer treatments, antibiotics, cholesterol-lowering medication, and improved drug delivery methods, much remains to be discovered [2]. One of the actual problems that has fundamental importance to personalized medicine, cancer diagnostics, and disease prevention is cell classification, since the accurate identification and characterization of cells based on their unique characteristics provides valuable insights into a patient's health status and help guide treatment decisions [3]. However, due to differences in shape and size, as well as some external environmental impacts, the precise classification of different cell types has remained a challenging task [3].

As a consequence of technological advancements, the fields of biology and biomedicine have experienced significant progress in the use of machine learning, which has become a valuable tool for tackling complicated tasks such as cell classification [3]. Machine learning techniques, such as deep learning, have demonstrated remarkable precision in a variety of tasks, such as object detection, semantic segmentation, and image captioning, and have been successfully applied to various biological domains, including regulatory genomics and electron microscopy [4]. In recent studies, there have been several attempts to develop cell classification based on machine learning techniques by training microscopic images. The studies that aimed to classify cells based on bright-field microscopy image datasets were unable to accurately reflect different types of cells [5]. Even though bright-field microscopes are effective at magnification, the low resolution of microscopes hinders the ability to discern the specific details of a magnified image [6]. Nevertheless, studies on application of deep learning methods based on fluorescence microscopy images have yielded optimal outcomes and are able to classify various cell types [7]. For example, studies conducted by Pärnamaa and Parts has achieved accurate classification of protein subcellular localization on fluorescence microscopy images of yeast cells [4,7]. Due to the use of fluorophore in fluorescence microscopy, it provides more comprehensive understanding of cellular morphology and functions, therefore fluorescence microscopy images have higher resolution than brightfield images [8] . The use of fluorescent dyes causes fluorescence microscopy to be more expensive technique than brightfield microscopy, hence the availability of fluorescence

microscopy image datasets is limited [9]. Considering the conclusion of recent studies, the concept of developing an alternative classification method for microscopy cell images into four cell treatments through the use of techniques of deep learning has emerged.

This thesis aims to explore the potential of deep learning to automate the classification of microscopy cell images into four cell treatments: Palbociclib, MLN8237, AZD1152, and CYC116. Each treatment is an anti-cancer drug, hence accurate and efficient automation of cell treatment classification through the application of deep learning has the potential to revolutionise cancer treatment and improve patient outcomes.

The additional objective of this thesis is to compare the potential of deep neural netoworks in automating the classification of fluorescent and brightfield microscopy images. Fluorescent microscopy images have a higher resolution than brightfield microscopy images, allowing for the detailed capture of cells; thus, fluorescent microscopy is a more expensive imaging technique [8,9]. Consequently, it is remarkable to compare the outcomes of the automated classification of fluorescent and brightfield microscopy images into four cell treatments.

This thesis is structured into five parts. Section 1 one of the most difficult tasks in biomedical research, known as cell classification, along with the necessary background information on microscopy image techniques for comprehending the results of the experiments carried out for this thesis. Section 2 covers the description and collection of raw data, as well as the preprocessing techniques applied to raw data. Section 3 outlines the methods used in the study, describing the classification algorithms, the computational environment and the general deep learning pipeline used in the experimental stage of this study. Section 4 desribes different experimental approaches and explains the results. Section 5 provides analysis of the results of the work, including the interpretation of the results of experiments. The findings of study is compared to the previous works's results and the possible future research approaches are disscused. The main concept and achievements are concluded.

1 LITERATURE REVIEW

This chapter provides an overview of one of the most difficult tasks in biomedical research, known as cell classification, along with the necessary background information on microscopy image techniques for comprehending the results of the experiments carried out for this thesis. In addition, this chapter examines recent studies conducted in the field of biomedical imaging that have contributed to the achievement and conclusions of study results.

1.1 Microscopy

Microscopy is the visual examination of invisibly small objects, such as cells, using the most powerful scientific instrument, the microscope. The microscope consists of lenses that magnify objects, and the vast majority of contemporary microscopes use additional lenses to increase magnification; these microscopes are known as compound microscopes [6]. White light-illuminated microscopes are brightfield microscopes, and they are effective at magnification, which makes objects appear larger. However, the low resolution of brightfield microscopes makes it difficult to observe the intricate details of a magnified image [6]. Fluorescent microscopes are frequently used to visualize specific characteristics of small samples [8]. The high resolution of fluorescence microscopes is achieved by labeling the sample of interest with a fluorescent substance known as a fluorophore and illuminating it through the lens with a high-energy source [8]. Fluorophores are organic molecules that are frequently charged and contain aromatic rings, making them toxic to organisms [10]. In comparison to brightfield microscopy, fluorescence microscopy is a more complex, time-consuming and costly technique due to the use of fluorescent dyes [9]. Other types of microscopy include phase-contrast, confocal, 4D live-cell imaging, and automated microscopy [11]. Phase-contrast microscopy displays intact cells, whereas confocal microscopy is used for rapidly picturing cells as a whole [11]. 4D live-cell imaging integrates confocal with time-lapse microscopy, and automated microscopy utilizes software features to effectively focus in each field of view to generate quantitative information [11].

The universality of microscopy enables researchers to employ a wide range of imaging techniques to explore and evaluate cell-based assays, which are vital tools for understanding the behavior of compounds in biological systems. The most prevalent cell-based assays include 3D cell models, cancer research, cardiomyocytes, cell counting, cell migration assay, cell painting, COVID-19 and infectious disease research, drug discovery and development, live cell imaging, neurite tracing, stem cell research, and toxicity screening [11].

Microscopy's end-to-end pipeline that has been notoriously challenging for a long time consists of three steps: image acquisition, image analysis and the creation of useful outputs [12]. Obtaining a high-quality image of an inspected item involves manual calibration of numerous hardware parameters, which is a time- and labor-intensive task requiring extensive domain-specific knowledge [12]. Once the microscope has produced a set of high-quality images, automation takes over to prepare for in-depth examination. Medical research and clinical activity rely on the detection and classification of microscope image contents [12]. In order to diagnose cancer, for instance, researchers must be able to distinguish between normal and cancerous cells. If performed manually, researchers would have to inspect and annotate multiple datasets with great care. When using an artificial intelligence (AI) model, however, researchers would only manually annotate one dataset with correct images [12]. The artificial intelligence model can distinguish between images in successive samples in seconds, a task that would have taken researchers hours to accomplish manually [12]. In the final step of the microscopy pipeline, analysis results are turned into scientific reports, which require a lot of precise data [12]. By automating image analysis, it is evident that AI could represent a step change in microscopy's capabilities.

1.2 Machine learning

Human intelligence is characterized by the capacity to apply prior knowledge to novel situations and to discern meaning in patterns. The primary objective of Artificial Intelligence (AI) is to replicate these abilities in non-human agents [13]. Machine learning is a subfield of Artificaial Intelligence (AI) that extracts features from large data sets and utilizes them to make predictions or decisions on unseen data [14,15]. The main goal of machine learning is to make classification or predictions, uncovering key insights in all disciplines by becoming the key of innovation in today's data-driven society.

The key elements of machine learning are:

- representation, which is the process of demonstrating knowledge;
- evaluation, which is the differentiation and assessment of machine learning;
- optimization, which is the generation of models [16].

1.2.1 Machine Learning methods

Machine learning can be divided into several categories depending on the type of data provided, the algorithm of training of data, and the evaluation of test data [17]. Machine learning models fall into three primary categories:

1. Supervised learning - machine learning algorithms that use labelled datasets to classify data or predict accurate outcomes [18]. The most prevalent example is image or object detection, in which the machine learning model classifies objects from videos or images, thereby making them applicable to numerous computer vision techniques and imagery analysis [19].
2. Unsupervised learning - machine learning algorithms that analyse and cluster unlabelled datasets by discovering hidden patterns or data groupings without the need of human intervention [18]. Anomaly detection, for instance, utilizes machine learning models to identify unusual data points in a dataset, such as fraudulent transactions [20].
3. Reinforcement learning - machine learning algorithms that are similar to supervised learning, but algorithms aren't trained using data, but model learns as it goes by trial and error [18]. This type of machine learning algorithms are very widely used in natural language processing (NLP) in text summarization, question answering, and machine translation [21].

Futhermore, there is a common approach to machine learning that seeks better predictions by combining predictions from multiple models, which is called ensemble learning [22].

Researchers face several challenges when imaging biological specimens. The machine learning technique called deep learning has the pontential to address some of these problems [13].

1.2.2 Deep Learning

Deep learning (DL) is a subset of machine learning that is based on learning and improving on its own by examining computer algorithms [23]. Deep learning uses artificial neural networks to enable algorithms to learn in a manner analogous to that of humans, whereas machine learning employs simpler concepts. Convolution Neural Network (CNN) is a class of deep neural network widely used for image processing and also inspired in biological process. Several studies revealed that in cell classification CNN performed well than human experts

[5]. Deep neural networks are widely used in variety of applications such as medicine, computer vision, etc [5].

1.2.3 Deep learning applications

Deep learning algorithms implement neural networks with multiple hidden layers to achieve high levels of pattern recognition. These hidden layers operate as the algorithm's memory, enabling it to remember and store training-observed patterns. The capacity to preserve patterns allows deep learning algorithms to make predictions on new data using the patterns that have been stored. This aspect of deep learning has led to significant advancements in numerous fields, such as image classification, object detection, and face recognition, natural language processing and autonomous driving [24].

1.2.4 Related work in image classification

Image classification is an essential task in computer vision, with the objective of classifying an image into one or more predetermined classes. Similar to the majority of other scientific disciplines, biomedical imaging has progressed due to advances in deep learning.

Shilman et al. investigate the application of deep learning algorithms to the classification of microtubule network images [7]. Microtubule networks within cells serve as indicators of the presence of chemical compounds. The research developed convolutional neural networks (CNN) using Keras and Tensorflow to classify images of microtubule networks into three classes, representing three degrees of chemical agent exposure [7]. For each class, 1,000 to 1,200 grayscale images of individual cells with a visible microtubule network were captured. In order to improve the results of CNN training, data augmentation techniques such as rotation and sharpening were applied to the original dataset. CNNs perform significantly better in the task of differentiating various levels of chemical agents, increasing classification accuracy from 52% for the best human expert to 70.5% [7]. The results of the study indicate that the deep neural network classifies cells better than human experts [7].

In another study, Pärnamaa and Parts developed a neural network called DeepYeast for classifying the subcellular localization of fluorescent proteins in microscopy images of yeast cells [4]. The dataset consisted of 7132 microscopy images divided into 12 classes, each representing a compartment of cell, such as cell periphery, cytoplasm, endosome, endoplasmic reticulum, Golgi, mitochondrion, nuclear periphery, nucleolus, nucleus, peroxisome, spindle pole, and vacuole [4]. Each image was divided into 64x64 pixel patches centered on the

midpoint of the cell, resulting in dataset of 90,000 cell images. The study's findings indicate that the DeepYeast neural network can achieve a classification accuracy of 91% for individual cells across 12 subcellular localizations (classes) and 100% for proteins when entire cell populations of at least moderate size are considered [4].

Deep learning has proven to be an excellent tool for image classification tasks, as demonstrated by the examples. Even if the previously described model accuracies are relatively high, classification models can be optimized and their performance can be enhanced. To improve the accuracy of high-performing models based on deep learning algorithms, it is preferable to employ large datatests.

2 DATASET

This chapter covers the description and collection of raw data, as well as the preprocessing techniques applied to raw data.

2.1 Fundamentals of image data

The data utilised in this study are images of cells. Each image consists of pixels, which are the smallest addressable element [25]. Each pixel is represented by a combination of three colours, namely Red, Green, and Blue, which is referred to as an RGB image [25,26]. Since each number is an 8-bit number, each colour value can range from 0 to 255, and if all three values are at maximum intensity, they equal 255, which displays as white. However, if the value of all three colours is 0, the colour displayed is black [26]. Since each value can have 256 different intensities or brightness levels, there are approximately 16.8 million (256^3) distinct colour shades [25].

A pixel matrix constitutes an image. Figure 1 depicts an example of an image of the number 8, which is a 3*5 (pixels) image in which each square represents a pixel and the computer addresses it as a 3*5 matrix with three numbers for red, green, and blue [25]. In addition, the number of primary colours (R, G, B) is referred to as channels, and in the field of machine learning, images are viewed as matrices on multiple channels [25]. Therefore, on a computer, the image on Figure 1 will be three 3*5 matrices if we consider it as RGB, however, if we only consider a grayscale image, then it is just one 3*5 matrix [25].

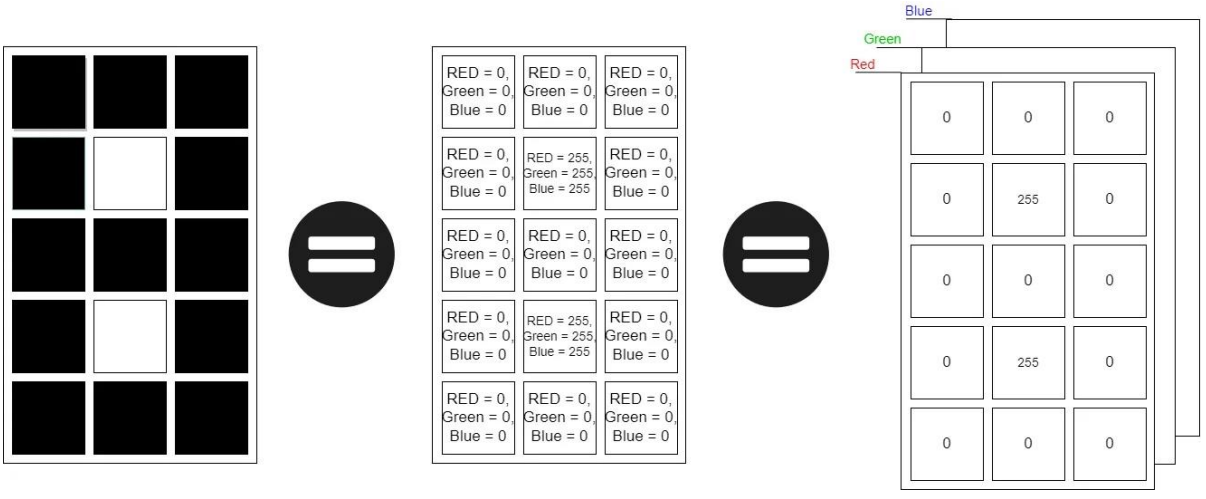

Figure 1: Illustration of an RGB image format [25]. Each image consists of a matrix of pixels, with each square representing a pixel and containing three numbers representing red, green, and blue.

2.2 Raw data collection

Under the terms of a cooperation agreement with the University of Tartu, the data used in this PerkinElmer supplied the data and supporting materials used for this thesis. The dataset contains images of cells and is referred to as cell painting assay data. Cells are painted with five fluorescent dyes to label different cell components, including the nucleus, endoplasmic reticulum, mitochondria, Golgi apparatus, and RNA [27]; the provided data is based on high-content images. By utilising the cell painting assay, it will be possible to determine the specific biological state of a cell, enabling the classification of cells treated with various drugs [27]. PerkinElmer utilises high-content microscopy to acquire this type of image-based data with a high level of detail. The primary data collection steps are depicted in Figure 2. In this thesis, deep learning is used to automate the classification of microscopy images into cell treatments for the purpose of data analysis.

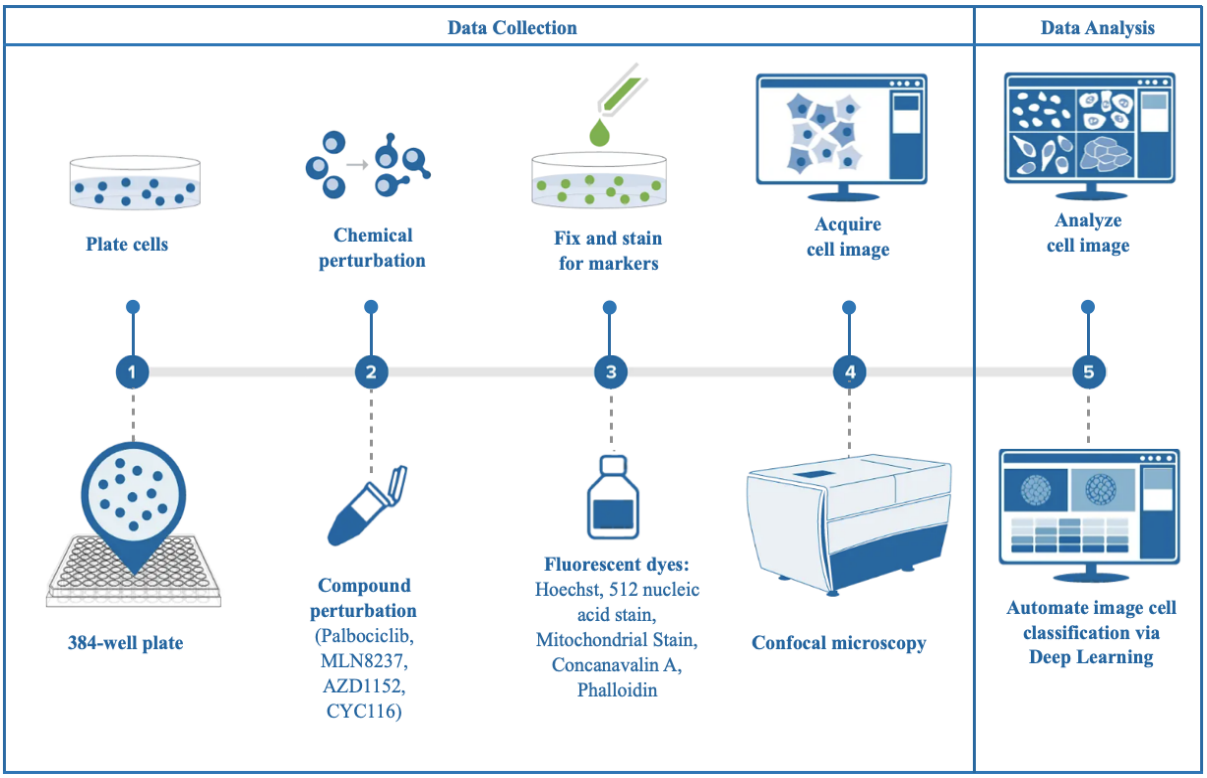

Figure 2. Representation of the data collection and analysis processes. PerkinElme collected data in five steps, including cell plating, compound perturbation, dye straining, and image acquisition. This work does the data analysis part, by automating the classification of microscopy images into cell treatments using deep learning.

2.3 Raw data description

Figure 3 illustrates the content of the microplate with 384 wells, and the provided data includes images from 16 rows of our 24-row microplate comprising 264 wells. The total dataset contains 1056 microplate images with a resolution of 1080 x 1080 from eight distinct categories of cells. Three classes are positive controls, one class is a negative control (DMSO), and four classes represent different cell treatments: Palbociclib, MLN8237, AZD1152, and CYC116. Given that the purpose of this study is to classify microscopy cell images from four cell treatments, the subset of data containing only images from these four cell treatments contains 656 cell images.

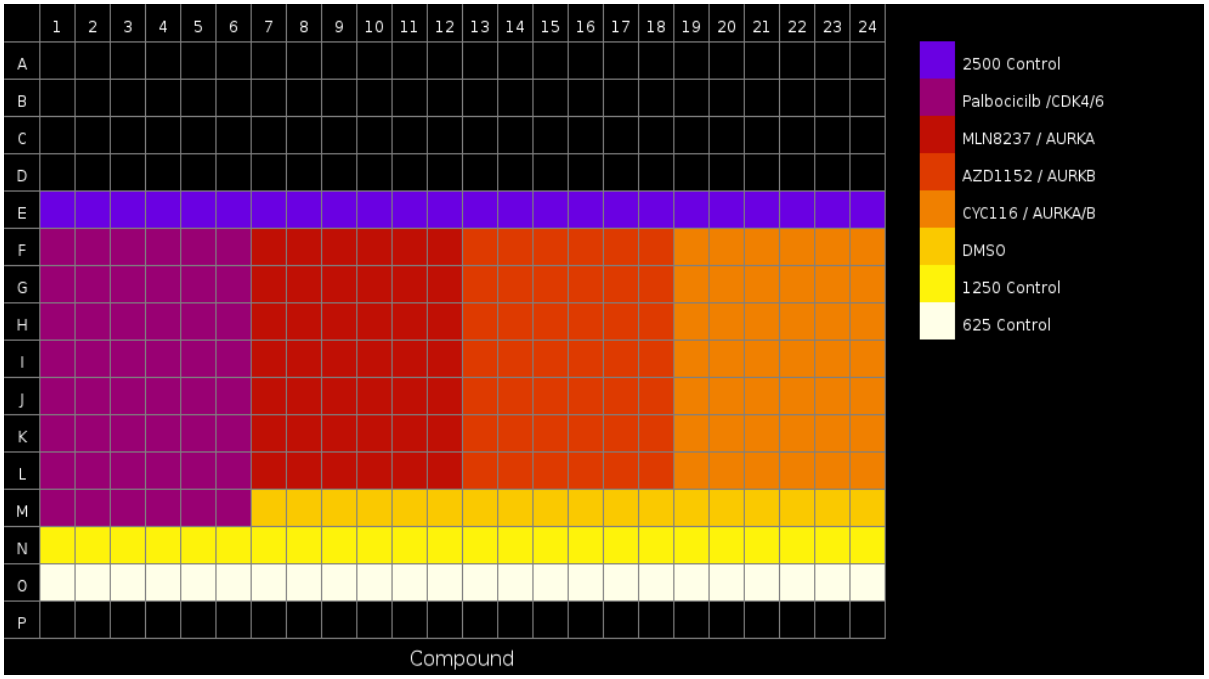

Figure 3: Representation of the content of PhenoPlate 384 microplate. Only 264 wells of 384 are used and contain content from eight distinct classes of cells. Three classes are positive controls, one class is a negative control (DMSO), and four classes represent different cell treatments: Palbociclib, MLN8237, AZD1152, and CYC116.

As mentioned in Section 2.1, the number of primary colours (RGB) is referred to as channels, and a typical RGB image contains three channels. Each image in the provided dataset consists of seven channels: five fluorescence channels and two brightfield channels of the same sample. Figure 4 depicts seven distinct image channels for each cell treatment.

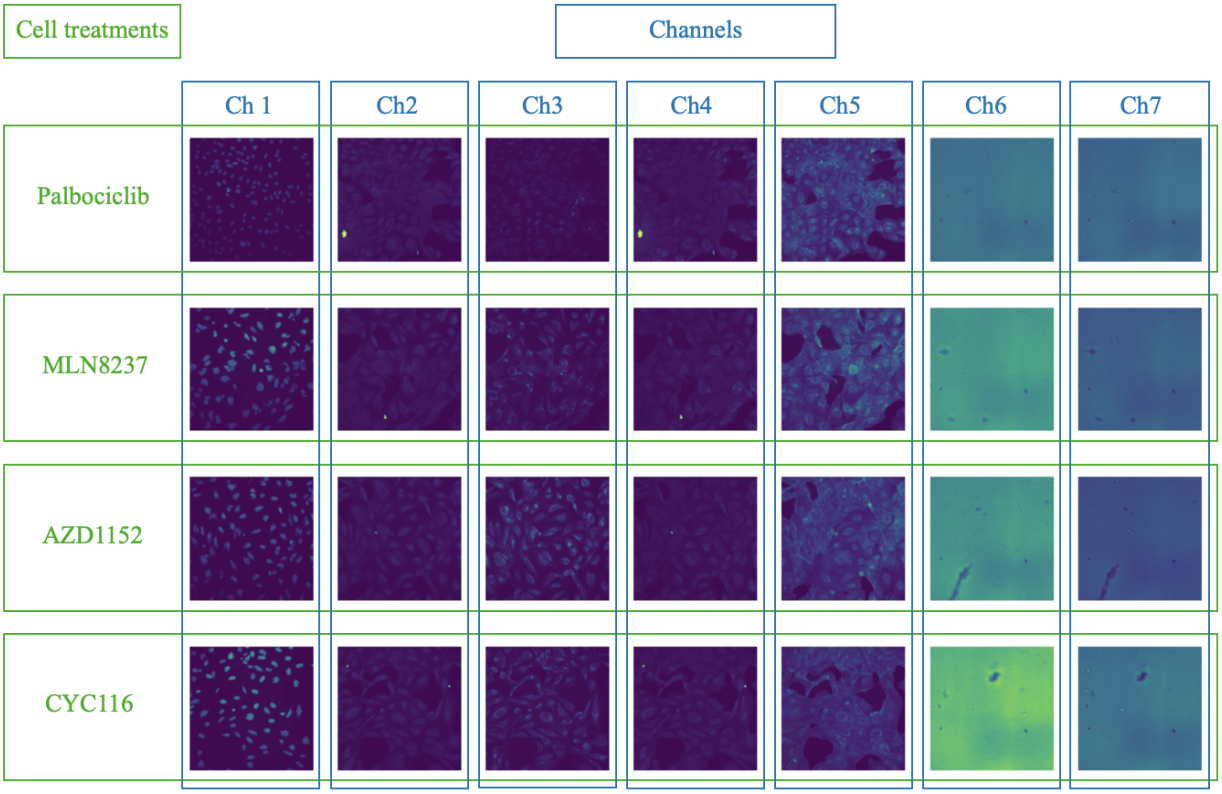

Figure 4. Example of the image of each cell treatment across seven image channels. Each row in green represents a cell treatment, while each column in blue represents an image channel. There are four cell treatments; and seven image channels, five of which are fluorescent and two of which are brightfield.

The first five channels of the image are fluorescence channels, and each fluorescence channel is stained with a different dye, each to label the corresponding cell component:

- in channel 1, PhenoVue Hoechst 33342 dye was used, which emits fluorescence upon binding double stranded DNA and stains the nuclei [28].
- in channel 2, PhenoVue 512 Nucleic Acid Stain dye was used to stain the RNA enriched organelles, such as nucleoli found in the nucleus of mammalian cell [29]
- in channel 3, PhenoVue 641 Mitochondrial Stain is a fluorescent dye which accumulates in heathly mitochondria of live cells [30]
- in channel 4, PhenoVue Fluor 488-Concanavalin A is a fluorescent lectin which displays high affinity for glycoproteins and glycolipids present at the cellular membranes, used for cellular membrane staining, particularly the endoplasmic reticulum [31]

- in channel 5, PhenoVue Fluor 555-WGA 568-Phalloidin is fluorescent lectin which displays high affinity for sialic acid and N-acetylglucosamine residues of glycoproteins and glycolipids present at the cellular plasma membranes. It can be used for cellular membrane staining, particularly the Golgi apparatus [32]

As shown in Figure 4, the last two channels of each image are brightfield channels: channel 6 and 7. These channels of the image have different focal distances, low and high, respectively.

2.4 Data preparation and pre-processing

Only images of 4 cell treatments from dataset were chosen for classification. As shown in Figure 3, the provided dataset of images of cell treatments such as Palbociclib, MLN8237, AZD1152, and CYC116 is quite balanced. There are 192 images of class Palbociclib, 168 images of class MLN8237, 168 images of class AZD1152, and 168 images of class CYC116, for a total of 696 images. All images in the dataset are stored as TIFF files, and each image was converted into a pixel array. The dataset was then divided into initial training, validation, and testing datasets with 60% of the images reserved for training, 20% for validation, and 20% for testing, respectively. These datasets were utilised throughout the entirety of the study.

Pre-processing

Both the training and testing datasets' images were preprocessed. Normalisation of images is a common preprocessing technique in machine learning and computer vision tasks. Normalisation adjusts an image's pixel values to a standard range. This will aid in maintaining numerical stability during the model's training process. Normalisation was applied differently to single-channel images, such as grayscale images, and three-channel images, such as RGB. The normalization method used for single-channel images is calculated using the following formula [33]:

$$IN = (I - Min) * \frac{newMax - newMin}{Max - Min} + newMin, \text{ where}$$

- IN is a array of normalized image pixels;
- I is the array of image pixels
- Min and Max are minimum and maximum pixel values of the image.
- $newMin$ and $newMax$ are new minimum and maximum pixel values of the normalized image. In this thesis experiments, $newMin$ and $newMax$ are set to 0 and 255, respectively.

The normalization method used for 3-channel images is called z-score normalization. Z-score normalization, also known as standardization, is a method of normalizing data by transforming it to have a mean of 0 and a standard deviation of 1 [34]. The process of z-score normalization involves subtracting the mean of the data from each data point and then dividing it by the standard deviation [34]. This transformation ensures that the resulting values have a distribution with a mean of 0 and a standard deviation of 1. Normalized images in dataset is calculated using the formula :

$$IN = (I - mean) / standard\ deviation, \text{ where}$$

- IN is array of normalized image pixels;
- I is the array of image pixels
- x is the original image.
- mean and standard deviation are calculated along the batch, height, and width dimensions.

Then, normalized 3-channel images were converted to the range of 0 to 255 by using the following steps:

1. We identified the minimum and maximum pixel values of the normalised image array.
2. By subtracting the minimum value, multiplying by 255.0, and dividing by the range, the values in the array are scaled to the desired range of 0 to 255.
3. To ensure that the scaled values fall within the valid pixel value range, they are rounded to the nearest integers.
4. The array is converted to 8-bit integers.

In all experiments conducted for this study, images were normalised using the methods described.

Division into patches

Patches are small subregions or sections of an image in the context of image processing and computer vision. These patches, which are extracted from the original image to analyse are typically square in shape. Patches can be utilised for a variety of image-related tasks, such as object detection, image classification, and image segmentation. By extracting patches from an image, it is possible to analyse the image at various scales or resolutions.

In this research, there are two classification architectures that are used as classifier, Residual Networks (ResNet) and Transformers, which are explained later in the Section 3.2.

Transformers incorporate the algorithm for dividing images into patches as part of their architecture; therefore, no additional pre-processing step is required for dividing images into patches. In order to feed the original images to the transformer model, the images are rescaled to the appropriate input shape of the model as part of the working pipeline of the transformer.

In contrast, when employing the ResNet model, the image dataset of 696 images with a resolution of 1080x1080 pixels is divided into 16 non-overlapping patches of equal size and then passed as input to the ResNet classification model. As a result, the size of input data will increase by a factor of 16, which can improve the performance of deep neural networks, as a larger dataset can produce more accurate models. Figure 5 depicts an example of an image that has been divided into patches.

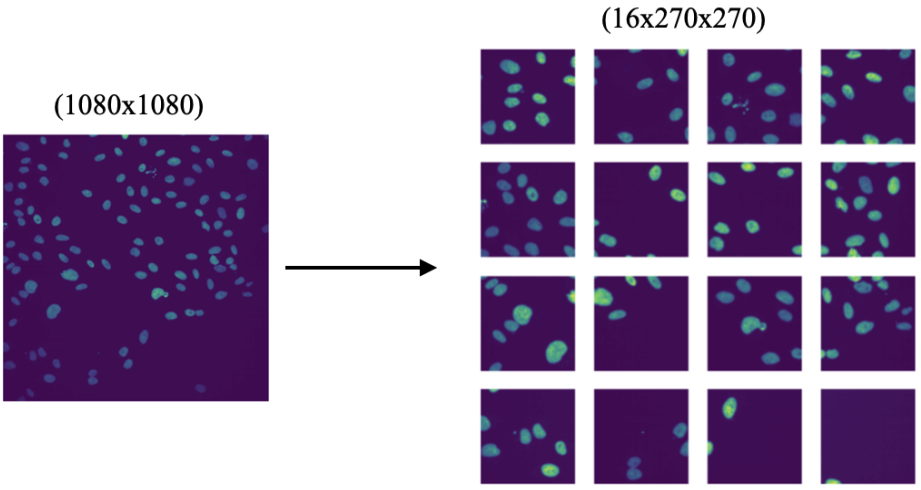

Figure 5. Representation of the process of diving a sample image with dimentions of 1080x1080 pixels into16 non-overlapping patches with dimensions of 270x270 pixels each.

3 METHODS

This chapter outlines the methods used in the study, describing the classification algorithms, the computational environment and the general deep learning pipeline used in the experimental stage of this study.

3.1 Deep learning software

The deep learning pipelines that are discussed in the thesis were implemented by us using Keras with Tensorflow as the backend and Pytorch. Keras is a high-level wrapper library that is open-source and designed to provide rapid experimentation with deep neural networks [35]. Keras was designed to speed up the process of training neural networks. Tensorflow is the backend framework that was used in this thesis; however, because Keras is flexible, it can use a variety of other backend frameworks to carry out the computation-intensive tasks. Tensorflow is a library that is primarily utilised for the purpose of conducting high-performance numerical computations. Its primary application is in the development of neural networks. Tensorflow, for instance, is able to make use of the cuDNN library to perform computation-intensive tasks on NVIDIA GPUs [36]. PyTorch, on the other hand, is a relatively new deep learning framework that is efficient in its use of memory, is flexible, and speeds up processing [35]. This thesis makes use of two different architectures, one of which is known as Residual Networks (ResNet), as well as two different Transformer models known as Vision Transformer and Swin Transformer. Pytorch library was utilised in the training of ResNet models, while Keras library with Tensorflow as the backend was utilised for training Transformer models.

Computation environment

Hardware used for this thesis is a computer with an access to University of Tartu's Graphical Processing Unit (GPU), which have high memory bandwidth to handle huge amounts of data, therefore allow for better processing of multiple computations simultaneously [37]. University of Tartu's High Performance Computing Center aims to develop the required infrastructure for scientific computing [38]. It provides 6 GPU nodes, the one used for the thesis has following hardware specifications [39]:

- 2 x Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz (48 cores total)
- 512 GB RAM
- 5TB of local SSD storage

- 24x NVIDIA Tesla V100 GPUs, with 16 GB of VRAM, only a single GPU was used at a time for training

For scientific computing of this thesis project, Jupyterlab was used. Jupyterlab is the web-based interface environment, which provides a flexible and powerful interface for creating, editing, and running code, as well as visualizing and analyzing data [40].

3.2 Classification architectures

Image classification is one of the most prominent applications of computer vision and belongs to the supervised learning subfield of machine learning [41]. Deep Learning is a technique that has been instrumental in the advancement of the computer vision field and has found widespread application in a variety of visual tasks including classification, segmentation, recognition, detection, and reconstruction. Deep Convolution Neural Networks (CNNs) have long been regarded as the method of choice when it comes to the classification of images, and the CNN models AlexNet, VGG, GoogleNet, and ResNet [41,42] are the most frequently used examples of this technique. Research [41,43] has shown that when compared to other models, the performance of ResNet models is significantly higher. Recently, Transformer-based architectures such as Vision Transformer (ViT) have achieved similar performance to ResNet architecture for image classification tasks [42]. In some cases, they have even surpassed it. However, certain aspects of the Transformer architecture, such as the application of non-overlapping patches, make one question the dependability of these networks [42]. As a result of this, the research makes use of both the ResNet and Transformer architectures. The results of comparing their respective performances can be found in the Section 4. Because of the limitations imposed by the format of the thesis, the artificial neural networks that were utilised are discussed in this section without going into extensive technical detail. The references to the pertinent literature are provided for each network to ensure additional information can be found.

3.2.1 ResNet architecture

ResNet is a Deep Learning architecure for image classification of Convolutional Neural Network (CNN), which is a class of deep neural networks, most commonly applied to analysing visual imagery. There are three pre-trained architurecture models: ResNet-34, ResNet-50 and ResNet-101, which are 34, 50, and 100 layers deep, respectively. The deeper the model, the stronger capability and the higher the classification accuracy [43].

ResNet stands for *Residual Network* and what makes it residual network is its identity connections. Identity connections takes the input directly to the end of each residual block, as shown below with the curved arrow in Figure 6 and Figure 7.

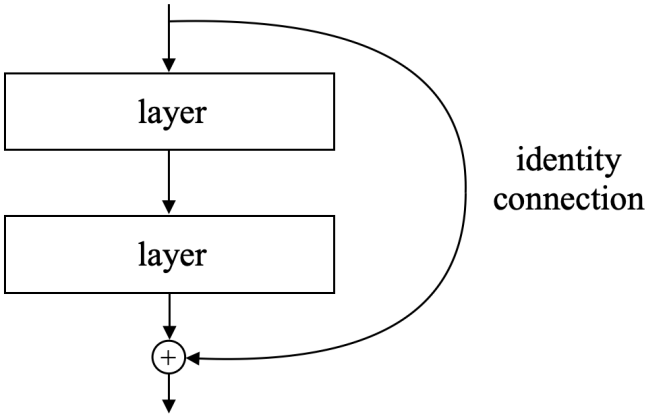

Figure 6. Representation of residual network. The curved arrow shows the identity connection.

Residual Network is a network with skip connections, in which the lower layer is directly connected to the upper layer via the shortcut connection that performs identity mappings, and the layer outputs are merged [41,44]. The identity skip connections, also known as residual connections, enable deep learning models with tens or hundreds of layers to be easily trained and to approach greater accuracy as the number of layers increases [44].

In this thesis research, pre-trained ResNet-50 architecture is used for all experiments, which was initially trained on a large dataset called ImageNet. The ResNet-50 has classical standard architecture, and no contributions were made to change the architecture itself. The architecure of ResNet-50 is 50 layers deep and each layer has its specific role in the arhitecture, the main components are convolutional layer, max pooling layer, residual block, average pooling layer and fully connected layer; the architecture is visualised in Figure 7.

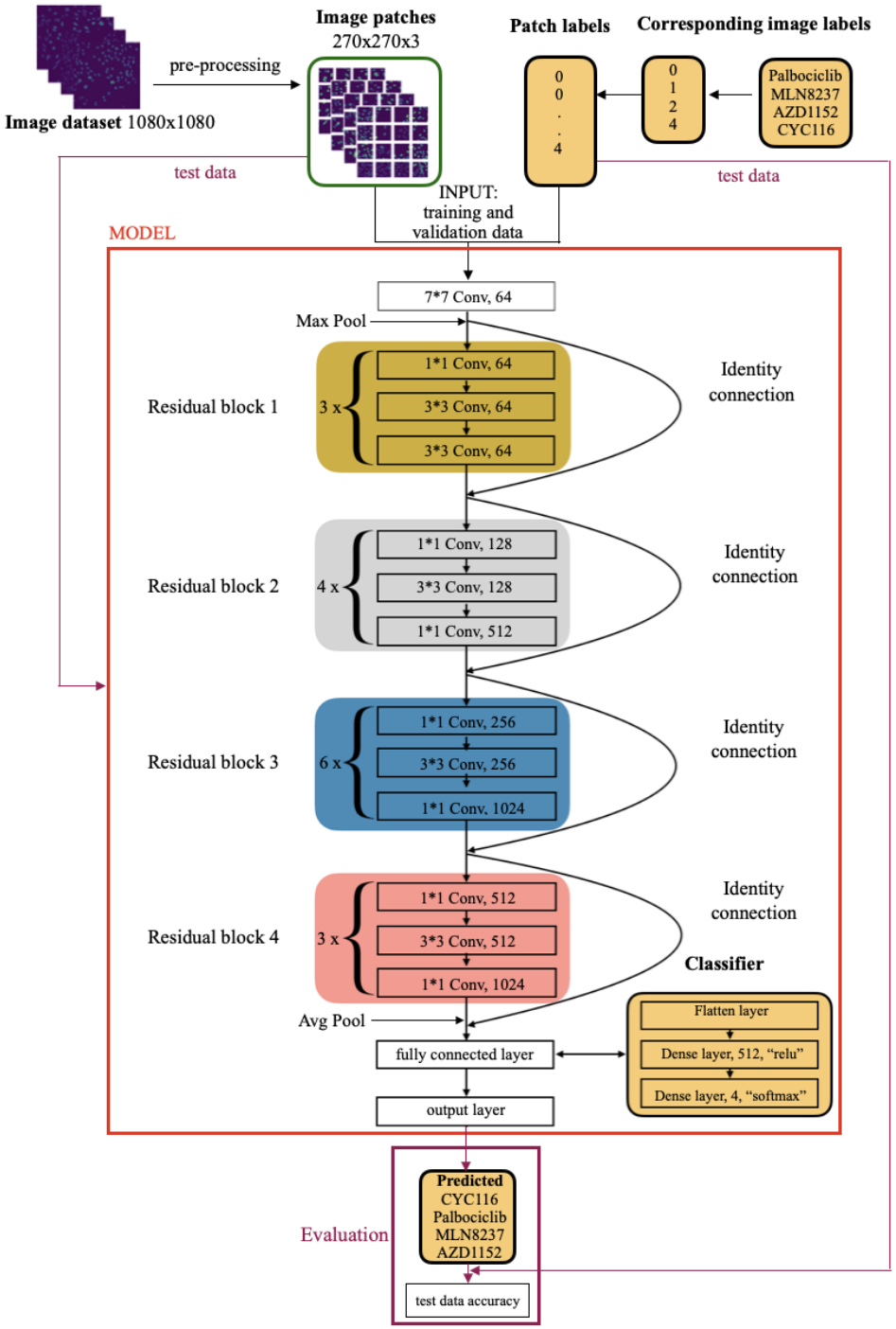

Figure 7. The figure shows the pre-trained ResNet-50 architecture. This also displays the ResNet model's working pipeline, including steps such as data loading, data pre-processing, the model interference, and the evaluation step.

The *convolutional layer* extracts features and patterns from the input layer by applying a convolution operation to the input images, which is done by a set of learnable filters called kernels that slide over the input data perform element-wise multiplication and summation.

The *Max Pooling layer* performs downsampling, where it selets the maximum value within the small region of input. By doing this, it reduces the size of feature maps while retaining the most important features.

The ResNet50 consists of 4 *residual blocks*, each containing multiple convolutional layers with shortcut connections as shown in Figure 6. The residual blocks are stacked together to increase the network's depth.

The *average pooling layer* uses spatial pooling to make the feature maps smaller by combining the information from different parts of the feature maps and calculating the average value.

The *fully connected layer* serve as ***classifier***, the specific details of this layer depends on the task. In this research experiments, the pre-trained model's fully connected layer is removed and few layers such as flatten and dense layers are added, in order to prodcue the final output probabilities for 4 classes.

3.2.2 Transformers

Transformers are a class of deep learning models that have been influential in various computer vision tasks such as image classification, object detection and image segmentation [45]. Transformer model consists of two main components: the encoder and the decoder [45]. The encoder procedures the input image and extracts high-level features, whereas the decoder uses these extracted features to perform specific tasks, such as classification [46].

The main idea behind transformers is self-attention, which allows the model to weigh the importance of different parts of image when processing it. Instead of using convolutional layers like CNNs, trasnformerws use self-attention mechanisms to capture global dependencies and relationships within the image.

To process an image, the input images are rescaled to the input shape of the transformer model, which divides an image into smaller regions called patches. Similarly described in Section 2.4, the preprocessing of input images being divided into patches will not be applied prior to the transformer model, since the transformer has this algorithm as part of its architecture to generate patches from input images, as visualised in Figure 8.

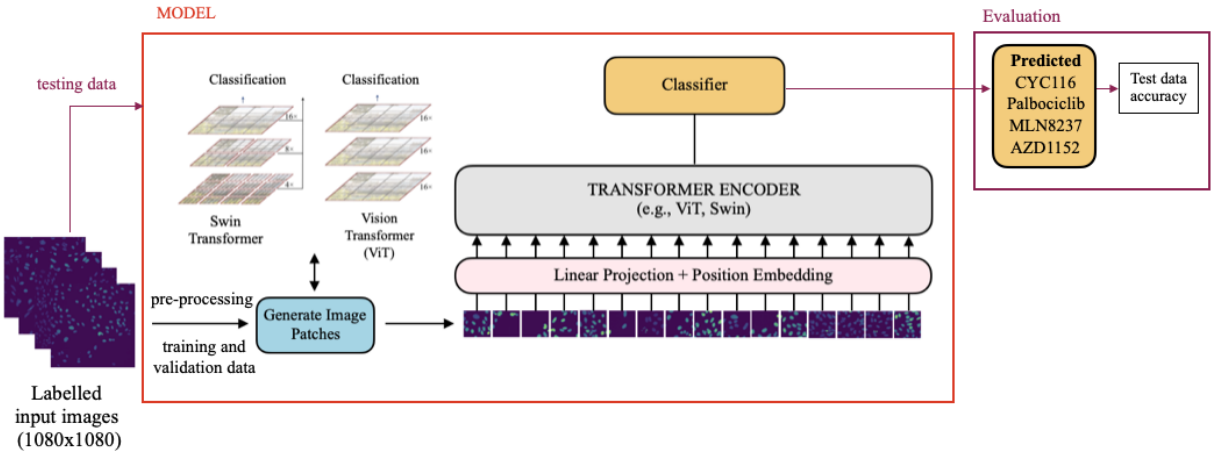

Figure 8: The figure shows the architecture of transformer model. This also displays the transformer model's working pipeline, including steps such as data loading, data pre-processing, the model interference, and the evaluation step.

Vision Transformers (ViTs) and Swin Transformers are two prominent types of transformers in computer vision. These architectures have shown remarkable performance in image recognition tasks such as image classification, each with its unique characteristics [45].

Vision Transformers (ViTs) process images by dividing them into smaller patches and linearly embedding each patch into a lower-dimensional space, with the which results patch embeddings supplied to a transformer encoder [47]. Self-attention layers and feed-forward neural networks make up the transformer encoder. Self-attention captures broad connections between patches, enabling the model to learn the relationships between various image components. Local patterns are captured by feed forward networks, which then apply nonlinear transformations. ViTs excel at modelling long-range dependencies in images, which allows them to identify objects or patterns which cover various parts. Nevertheless, when dealing with high-resolution images, ViTs may face computational and memory limitations. The pre-trained Vision Transformer model used in this thesis research is "google/vit-large-patch16-224-in21k", which is developed by Google [48]. This trasnformer model is commonly used for image classification tasks, and it was pre-trained using a large-scale dataset called "imagenet21k," which contains 21.000 classes of variety of objects [48]. This model divides input images with a resolution of 224x224 into 16x16 pixel patches. Architecture and pretraining of the model make it a potent instrument for accurate image classification.

Swin Transformers, on the other hand, offer an alternative solution to the limitations of ViTs. Swin Transformers employ a hierarchical approach to process high-resolution images

effectively [49]. The image is divided into smaller, non-overlapping patches, which are then hierarchically processed, which allows the model to effectively capture both local and global data. Swin Transformers are comprised of two distinct types of transformer blocks: window-based self-attention and shift-based window partitioning. The window-based self-attention blocks capture long-range dependencies within a window of a fixed size, whereas the shift-based window partitioning captures contextual information across multiple windows [49]. Swin Transformers achieve computational efficiency and memory scalability for high-resolution images by introducing window-based processing. This thesis employs the "microsoft/swin-large-patch4-window7-224", the pre-trained swin transformer model developed by Microsoft [50]. With an input size of 224x224 pixels, a patch size of 4, and a window size of 7, the model has been trained to excel at a variety of computer vision tasks, involving image classification, object detection, and semantic segmentation [50]. Overall, transformers leverage self-attention mechanisms to extract meaningful features from 3-channel images leading to more accurate predictions.

3.3 Training

The purpose of training a classification model is to teach it accurately classify input data into predefined classes. The goal is to enable the model to discover patterns within the data that will aid it to make accurate predictions regarding the class labels of unseen (test) data.

Throughout the training process, the model is presented with a labeled dataset, where each image is associated with a known class label, whcih is a type of cell treatment in our case. By doing this, the model learns from this labelled data by adjusting its internal parameters or pre-trained weights in accordance with the input features and their associated class labels. The training process involves minimizing a loss function that measures the discrepancy between the predicted class labels and the true class labels in the training data.

By iteratively updating the model's parameters using an optimization algorithm, the model's ability to generalise from training data to unseen examples is gradually enhanced. The optimization algorithm used is Adam optimizer. The trained classification model, optimized using the Adam optimizer, can then be used to classify new, unseen instances by applying the learned knowledge and making predictions based on the patterns it has learned during training.

The training of model is stopped by two scenarios. The first case, once the network has reached its smallest error data and the loss function on validation data doesn’t change any

more. In the second case, once the training lose is getting smaller than validation loss, leading the model to overfitting, which can decrease further model's perforance on unseen (test) data.

Moreover, ResNet and Transformer models were trained on seperate frameworks. The pre-trained ResNet50 model was trained on Keras with Tensorflow. The pre-trained Vision and Swin Trasnformer models were trained on Pytorch.

3.4 Evaluation

The evaluation step is essential for receiving assessment of how well the trained classifier model performed and optimising the parameters accordingly. After the classification model has been trained on train data, then the model is evaluated on unseen test data and as a result, probabilities for predicting each class are obtained. There are different evaluation metrics suitable for classification. The evaluation metrics used to measure the performance of models for futher comparison of models' performances. Accuracy is the most common method to measure the frequncy at whcih the classifier makes correct prediction. It is typically calculated as the percentage of correctly classified samples out of the total number of samples in the dataset. By evaluating model on unseen data, it calculates the accuracy via the following formula:

$$Accuracy = (Number\ of\ correctly\ classified\ samples) / (Total\ number\ of\ samples)$$

This formula provides the overall accuracy of the model's predictions across all classes. Despite the fact that the same evaluation metric, accuracy, is used to evaluate test data for ResNet and Transformers, the steps of evaluation for each architectue are distinct. Since the input images for ResNet50 model are preprocessed images, which 16 non-overlapping patches per image. When the model is trained on patches and evaluated on test image patches, the probabilitities for predictions are for each individual patch. In order to evaluate the accuracy of test images based on its patches, post-processing step is done. The post-processing includes appying of the the major voting algorithm to class labels of the patches. Major voting algorithm operates the way that, the mode class label of 16 non-overlapping patches of the image, become the image class label. After the predicted lables of images are gathered, it will be compared to the true labels and the overall image accuracy will be calculated via the abovementioned formula. Th post-processing step and evaluation of ResNet-50 model on test data is demonstrated in Figure 9.

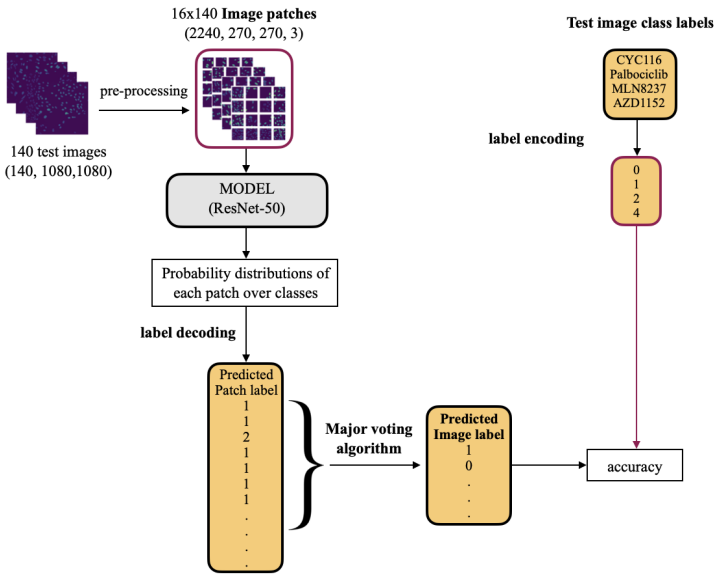

Figure 9. This figure visualises the working pipeline of the ResNet-50 model with additional post-processing step called Major Voting. Major voting algorithm operates the way that class label of the image is the mode class of its predicted (16) patch labels.

Transformers, on the other hand, contain the algorithm of division into patches in its architecure, therefore the pre-processing part will consisted of the rescaling the original images into the suitable input shape of the model. In case of ViT and Swin transformer models, the input shape of the images is 224x224. The evaluation of the transformer models on test images does not require any post-processing step to be done beforegand. Figure 10 below indicates the evaluation step of the tranformer model pipeline.

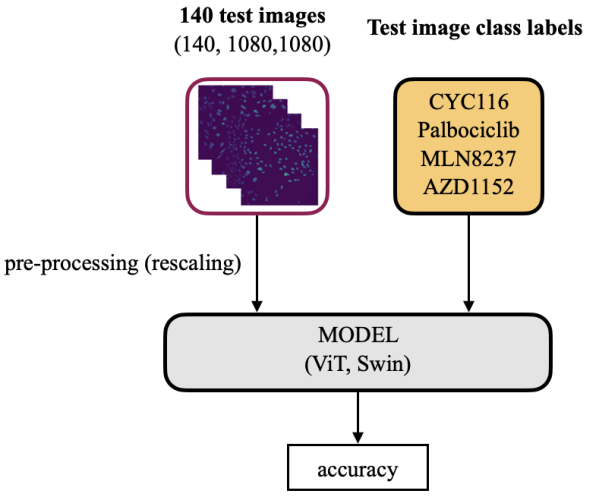

Figure 10. This figure visualises the general pipeline of transformer model, including the evaluation step.

3.5 Deep Learning (DL) model pipeline

"pipeline" is a series of steps or processes that are carried out in a specific order to transform and process the input data in the context of deep learning models. For the purpose of this study, a deep learning pipeline was utilised. This pipeline included data loading, data preprocessing, model inference, post-processing (for the ResNet model), and evaluation steps. The deep learning pipeline that was utilised for this thesis is illustrated in Figure 11.

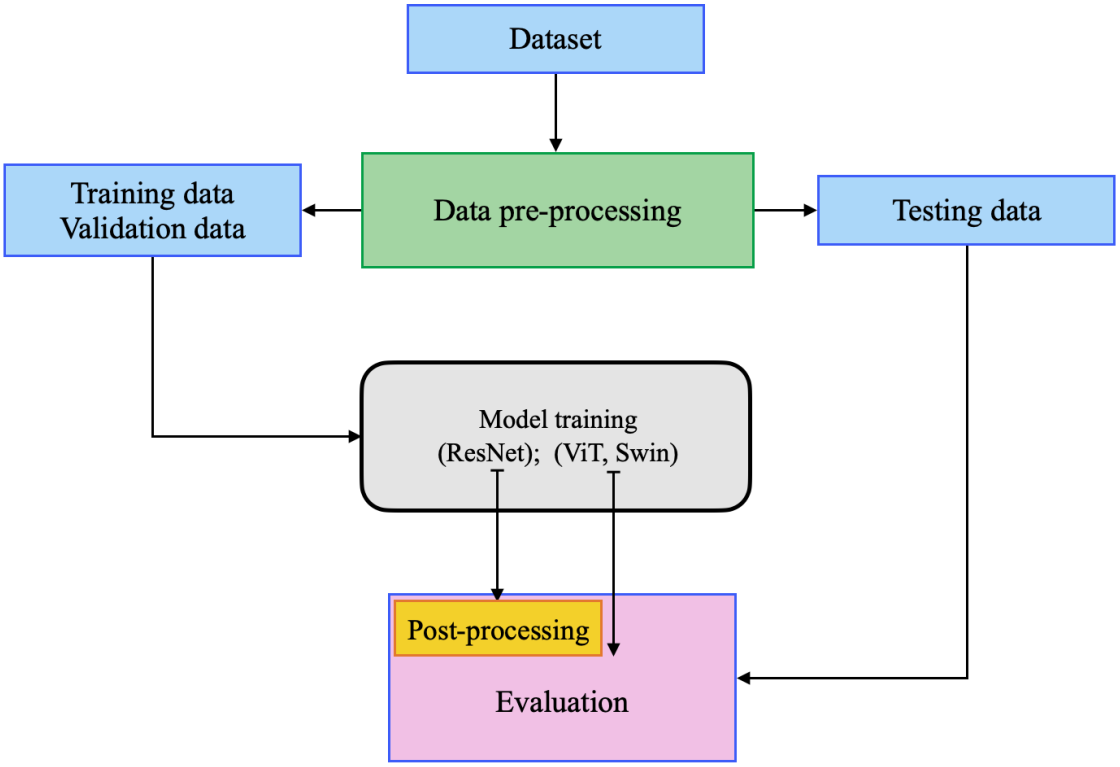

Figure 11: This figure displays the pipeline for deep learning models used in this study. This pipeline included data loading, data preprocessing, model inference, post-processing (for the ResNet model), and evaluation steps.

3.6 Supporting resourses

In order to enhance the clarity and coherence of the text, certain sections, particularly the introduction and methodology, have been revised, edited, paraphrased via the use of ChatGPT.

4 EXPERIMENTS AND RESULTS

This thesis aims to explore the potential of deep learning to automate the phenotype-based classification of microscopy images into four cell treatments: Palbociclib, MLN8237, AZD1152, CYC116. The additional objective is to compare the potential of deep neural netoworks in automating the classification of fluorescent and brightfield microscopy images.

As discussed earlier in Section 2.3, each image of the dataset consists of seven channels: five fluorescence channels and two brightfield channels of the same sample image. There were three distinct experimental approaches developed for the purpose of achieving the goal of this research. The first strategy that we tried consisted of applying deep learning models to the task of classifying cell treatments on each image channel independently. This method is also known as the single-channel classification of cell treatments. Analysing the results of our initial strategy's baseline performance, we have decided to use deep learning models to classify cell treatments on three image channels. This classification method is known as 3-channel classification of cell treatments. And our final approach consisted of using deep learning models to classify cell treatments on multiple image channels. This methodology is referred to as multi-channel classification of cell treatments.

4.1 Single channel classification of cell treatments

ResNet-50 model, ViT and Swin transformer models were utilized to classify single-channel microscopic images of cell treatments for all 7 channels of the image separately including 5 flourescent and 2 brightfield channels. Models were trained and evaluated in the same way for each channel and the entire workflow is visualised in Figure 11.

Since pre-trained ResNet-50, Vit, and Swin transformer models require input with 3 channels, we have performed an additional pre-processing step consisting of duplicating single-channel input twice in order to produce the 3-channel input. To perform this additional pre-processing step, python libraries such as OpenCV and Python Imaging Library (PIL) were used for ResNet and transformers, respectively. The process of preparing the input for pre-trained models is visualised in Figure 12.

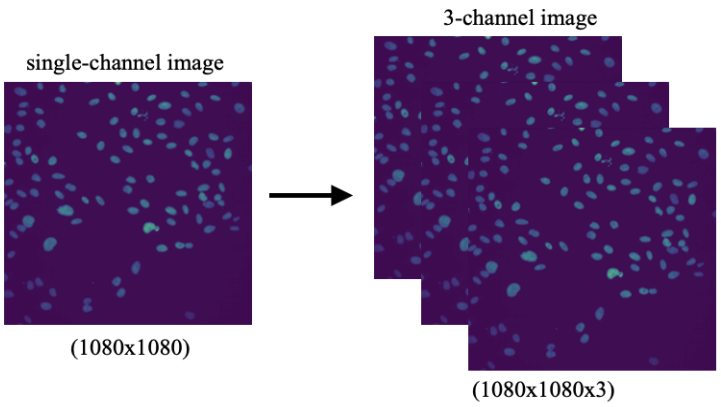

Figure 12. The additional pre-processing step consists of duplicating a single-channel input twice in order to generate the 3-channel input required by pre-trained models.

After training deep learning models to classify cell treatments on each image channel independently, we have obtained the results of this method, which are model evaluation accuracies based on test data. Table 1 displays the results of this experimental investigation.

Channel	**Fluorescent**					**Brightfield**	
	1	**2**	**3**	**4**	**5**	**6**	**7**
Model	**Accuracy on test image data, in %**						
ResNet	80.0	75.0	69.0	77.0	77.0	59.0	61.0
ViT	78.0	34.0	27.0	39.0	27.0	27.0	27.0
Swin	79.0	48.0	26.0	44.0	35.0	27.0	27.0

Table 1. The table displays the results of the initial classification strategy for cell treatments on each image channel separately. There are seven image channels, and the accuracies obtained by evaluating each model against test data are expressed in percentages (%).

Based on Table 1, it can be seen that the ResNet-50 architecture has performed better by reaching the highest test data accuracy on each separate channel. It is also noticable that all three models have achieved similar performance on channel 1. Interpreting the results from the perspective of channels, all three models which were trained on fluorescent channels have achieved higher accuracies, compared to models trained on brightfield channels. ResNet-50 model trained on channel 7 achieved slightly better accuracy than channel 6, reaching 61% and 59%, respectively, while the accuracies of transformers trained on channels 6 and 7 are the same and lower than ResNet's performance. This results demonstrate the superiority of the

ResNet model over more powerful models such as Vision Transformer and Swin Transformer. The reason for this is likely to be in the amount of the data available for the transfomer models.

4.2 Classification of cell treatments using multi-channel (3) image

Analysing the results of our initial strategy's baseline performance, we have decided to apply deep learning models to classify cell treatments on three image channels. In our initial apprach, when models were trained on single-channel images separately, three channels with the highest accuracies on observed validation data are chosen to be used in this second apporach. It is important to mention that the observed accuracies of validation data is reflected in the accuracies on test data shown in Table 1.

According to the results of the previous experiment the three fluorescent channels that have achieved highest accuracy are channels 1, 4, and 5. The process of combining separate single-channel fluoresecnet image into 3-channel fluorecent image is depicted in Figure 13.

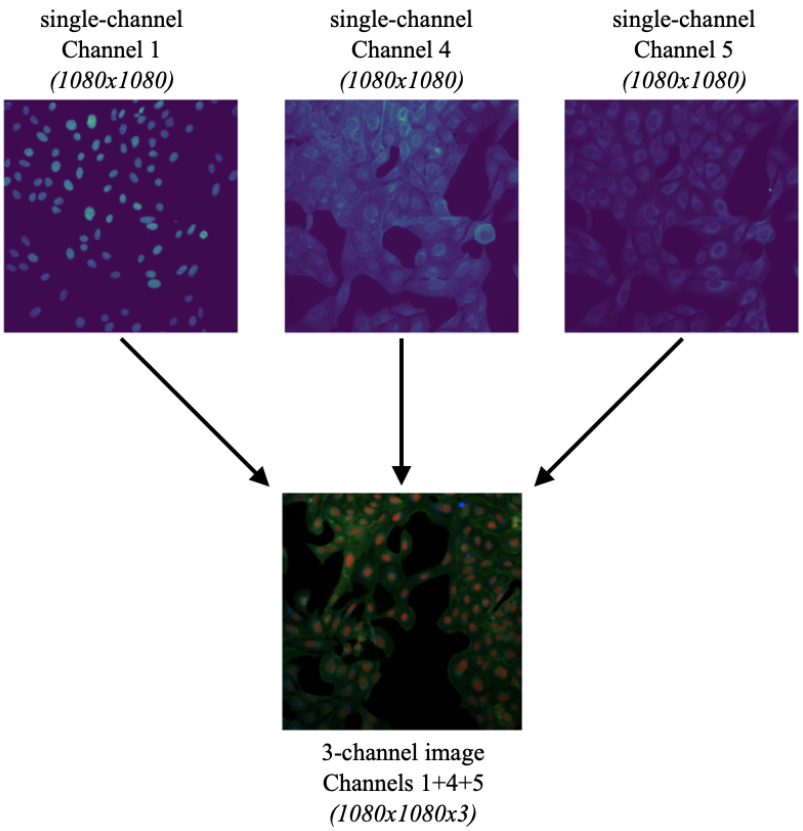

Figure 13: Example of combining three single-channel fluorescent images into one three-channel input image. Single-channel image of channels 1, 4 and 5 are combined to three-channel input image.

In the case of brightfield channels, there are only two channels: channels 6 and 7. To create three-channel images from two brightfield channels, one channel is duplicated. According to observed validation data accuracies that are similarly reflected in test data accuracies, models trained on channel 7 have achieved higher accuracy than on channel 6. Consequently, channel 7 represents two channels of the final 3-channel images, whereas channel 6 represents a single channel. Figure 14 depicts the method of combining two brightfield channel images into a three-channel input image.

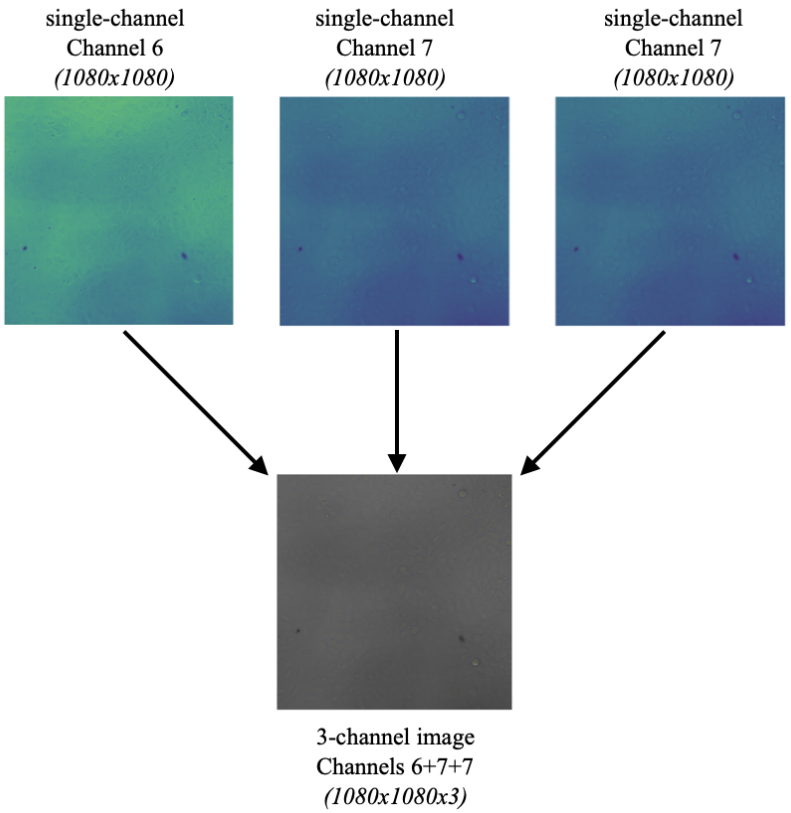

Figure 14: Example of combining two brightfield channel images into a three-channel input image. Channel 6 image represents one channel of the 3-channel output image and channel 7 is duplicated to represent two last channels of the 3-channel ouput image.

All three pre-trained models ResNet-50, ViT, and Swin were trained on multi-channel fluorescent and brightfield images and evaluated on testing data, which produced the results shown in Table 2.

Channel	Fluorescent	Brightfield
	1+4+5	6+7+7
Model	**Accuracy on test image data, in %**	
ResNet	84.0	52.0
ViT	38.0	27.0
Swin	86.0	59.0

Table 2. The table displays the results of the secondary classification approach for cell treatments on 3-channel fluorescent and brigthfield images. There are accuracies expressed in percentages (%) obtained from evaluation on test data.

According to the results of the experiments conducted using the second apporach, we have achieved higher accuracy by employing deep learning models to train on 3-channel fluorescent images to classify cell treatments. The concept of trainning deep learning models on 3-channel images that were combined from three channels single-channel images with highest accuracy, aimed to achieve better performance of classifcaiton of cell treatments, since multi-channel image consists of more information than a single-channel image. The results of this experimental approach demonstrate that the Swin transformer has the highest accuracy on both fluorescnet and brightfield images reaching the accuracy of 86% and 59% on test data, compared to the other two models.

Both, ResNet50 and Swin transformer models have improved their performance of classification of cell treatments on three-channel fluorescent images, compared to single-channel fluorescent images, while Vision Transformer model have fluctating accuracies across single-channel fluoresecnt images and not reaching the highest accuracy on 3-channel images.

Nevertheless, the performance is different in case of training deep learning models to classify cell treatments on 3-channel and single-channel brightfield images. The ResNet-50 model has achieved accuracies of 59% and 61% on two sing-channel brightifeld images, respectively, and have achieved accuracy of 52% on 3-channel brightfield images. The reason why ResNet-50 has increased its accuracy on multi-channel fluorescent images, compared to single-channel fluorescent images, but the perfomance worsened in case of brightfield images, could be because of the fact that fluorescent images have higher resolution than brightfield images. However, this still remains a mystery to me, because from the logical perspective, if

on seperate single-channel images, model has performed 59% and 61%, then on multi-channel images that are combination of these single-channel images, I would except to have at least the lowest accuracy of the channels, 59%. However, the accuracy achieved is equal to 52%. This could be a topic for future work and further studies focusing on the improvement on classificaion on brightfield images. The performance of the Vision Transformers on multi-channels brightfield images have not increased from performance on single-channel brightfield images, accuracy being equal to 27%.

4.3 Classification of cell treatments on several-channel (5) images

Since the results of the second experimental approach has demonstrated, that deep learning models have achieved higher accuracy when they are train on 3-channel fluorescent images to classify cell treatments, compared to the performance received on single-channel fluoresecent images. This has arisen the idea for the final approach to this task, which about using deep learning models to classify cell treatments several image channels, also referred to as several-channel classification of cell treatments. Since there are 5 fluorescent channels, the aproach will explore whether the performance of the same deep learning models can become better when trained on several-channel images, in this case, 5. Taking into account thelimitations imposed by the format and of the thesis, and longer expected time to finish this approach, we have been able to train only ResNet model on several-channel fluoresenct images so far. The reason for this is that the all three deep learning pre-trained models employed in this thesis require 3-channel input images, and in order to train deep learning model with 5-channel input images, need several modifications in arhcitecure of models. The ResNet-50 model trained on 5-channel input images have achieved test data accuracy of 65%, which is lower than the accuracy achieved on 3-channel input. The possible reason is since the ResNet model's architecture has been modified and initialised, and the weights used for training 3-channel images could not be used for training the input with 5-channel images, which has resulted in lower accuracy test data accuracy. Because of the limitations imposed by the format of the thesis and the expected time to finish this apporach is long, the training other two models with several-channel fluorescent images become necessity for future research.

5 DISCUSSION

This chapter provides analysis of the results of the work. Moreover, the interpretation of the results of experiments with explanation is covered. The findings of study is compared to the previous works's results and the possible future research approaches are disscused. The main concept and achievements are concluded.

The results of the initial and secondary classification approaches are combined and visualed in Table 3.

Channel	**Fluorescent**						**Brightfield**		
	1	**2**	**3**	**4**	**5**	**1+4+5**	**6**	**7**	**6+7+7**
Experimental approach	**1st**					**2nd**	**1st**		**2nd**
Model	**Accuracy on test image data, in %**								
ResNet	80.0	75.0	69.0	77.0	77.0	84.0	59.0	61.0	52.0
ViT	78.0	34.0	27.0	39.0	27.0	38.0	27.0	27.0	27.0
Swin	79.0	48.0	26.0	44.0	35.0	**86.0**	27.0	27.0	**59.0**

Table 3. The table displays the results of the initial and secondary classification approaches for cell treatments on single-channel and 3-channel fluorescent and brigthfield images. There are accuracies expressed in percentages (%) obtained from evaluation on test data.

It is the known fact that the fluorescent images have higher resolution than brightfield microscopy images due to the use fluorescent dyes, which allows for the detailed capture of cells; however, making it more expensive than brighfield microscopy [8,9]. Therefore, it is remarkable compare between the results of automated classification of fluorescent and brigtfield microscopy images into four cell treatments. The results of our initial strategy have demonstrated that, all three models have shown better performace on single-channel fluorescent images than on brightfield images, reaching the highest accuracy of 80% and 61%, respectively, both performed by pre-trained ResNet-50 models. Similarly, the results of our secondary strategy have demonstrated that, all three models have performered better on 3-channel fluorescent images than on brigtfield images, reaching the highest accuracy of 86% an 59%, respectively, both performed by Swin transformer model.

Image-based phenotyping of disaggregated cells using deep learning has been the subject of recent research [51]. The study classified eight distinct cell classes with a classification accuracy of 96.3% on 3-channel fluorescent images and 87.3% and 90.2% on individual channels [51]. The brightfield channel achieved the expected final accuracy of 37.7% [51]. When comparing the results of the most recent study with the results of this thesis, two similar patterns emerge. On 3-channel fluorescent images, the classification model has performed better than on single-channel fluorescent images, whereas the classification accuracy of brightfield channel images is lower than that of fluorescent channel images.

In conclusion, the results of three experimental approaches indicate that all three classifier models achieve greater accuracy on 3-channel fluorescent images than on single-channel fluorescent images. On 3-channel fluorescence images, the Swin Transformer model has achieved an accuracy of 86%. The Vision Transformer has generally demonstrated the lowest performance of the three models across all channels and experimental methods. The Swin Transformer has also achieved the highest level of accuracy, 59%, on multi-channel brightfield images, while ResNet has achieved the same level of accuracy on one of the single-channel brightfield images. According to the current results of third experimental approach, The ResNet-50 model trained on 5-channel input images have achieved test data accuracy of 65%. Since the not all results of this approach are gathered, the training other two models on several-channel fluorescent images become necessity for future research.

In general, classification models on 3-channel images have performed better than classification models on single-channel images, and classification models on fluorescent images have achieved the highest levels of accuracy compared to classification models on brightfield images.

SUMMARY

In this study, we explored the potential of deep learning to automate the phenotype-based classification of microscopy images into four cell treatments: Palbociclib, MLN8237, AZD1152, and CYC116. We investigated the capability of three pre-trained state-of-the-art deep learning models, such as ResNet, ViT, and Swin Transformer, to automatically classify brightfield and fluorescent microscopy images across single and multiple(3) channels. The corresponding results from single, multi-channel as well as from brightfield and floursecnet models have been comared and analysed.

The results reveal that the pre-trained ResNet-50 model outperforms the transformer models in the classification of single-channel fluorescent and brightfield images of microscopic cell treratments, achieving the highest accuracy of 80% and 61%, respectively. However, the Swin transformer performed better than the other models for cell treatment classification of multi-channel fluorescent and brightfield images, achieving the highest accuracy of 86% and 59%, correspondingly. Although experiments on multi-channel fluorescent images have resulted in higher levels of accuracy than on single-channel fluorescent images, the performance of brightfield images is rather different. Even though the highest accuracy on separate single-channel brightfield images was 61%, we were unable to obtain better performance from any model on multi-channel brightfield images, where the Swin Transformer model achieved only 59% accuracy. The results are quite surprsing, especially, provided previous research has shown that combining multiple channels yield better performance. This necessitates further investigation into the capacity of deep learning models for automating the phenotype-based classification of single- and multi-channel brightfield microscopy images. Although fluorescence microscopy allows for the detailed capture of cells, resulting in fluorescent images with higher resolution than brightfield images, fluorescence microscopy is significantly more expensive than brightfield microscopy. In addition, the use of certain fluorescent dyes in fluorescence microscopy may be toxic to organisms, which could result in adverse effects if used in high concentrations or improperly. Future research will examine the performance of deep learning models on cell treatment classification tasks based on microscopic brightfied channel images, with the goal of achieving the highest level of accuracy, as well as exploring the potential of these models to outperform fluorescent channel images. We hope that our work will inspire other researchers to gather and analyse different large biological image datasets in support of this initiative.

REFERENCES

1. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Visualizing Cells. Molecular Biology of the Cell 4th edition. Garland Science; 2002. Available: https://www.ncbi.nlm.nih.gov/books/NBK21048/

2. National Institute of General Medical Sciences. In: National Institute of General Medical Sciences (NIGMS) [Internet]. [cited 17 May 2023]. Available: https://nigms.nih.gov/

3. Oei RW, Hou G, Liu F, Zhong J, Zhang J, An Z, et al. Convolutional neural network for cell classification using microscope images of intracellular actin networks. PLoS One. 2019;14: e0213626. doi:10.1371/journal.pone.0213626

4. Pärnamaa T, Parts L. Accurate Classification of Protein Subcellular Localization from High-Throughput Microscopy Images Using Deep Learning. G3 Genes|Genomes|Genetics. 2017;7: 1385–1392. doi:10.1534/g3.116.033654

5. CELL CLASSIFICATION IN MACHINE LEARNING. In: MSRF|NGO [Internet]. 26 Aug 2021 [cited 17 May 2023]. Available: https://www.madrasresearch.org/post/cell-classification-in-machine-learning

6. Website. Available: https://www.thermofisher.com/ee/en/home/life-science/cell-analysis/cell-analysis-learning-center/molecular-probes-school-of-fluorescence/fundamentals-of-fluorescence-microscopy/how-fluorescence-microscopy-works.html

7. Deep Learning of Cell Classification Using Microscope Images of Intracellular Microtubule Networks. [cited 17 May 2023]. Available: https://ieeexplore.ieee.org/document/8260606

8. Fluorescent. In: Microscopy [Internet]. 2007 [cited 17 May 2023]. Available: https://serc.carleton.edu/microbelife/research_methods/microscopy/fluromic.html

9. 16 Difference Between Brightfield And Fluorescence Microscope. In: Microbiology Note – Online Biology Notes [Internet]. Sourav Bio; 21 Jul 2020 [cited 17 May 2023]. Available: https://microbiologynote.com/16-difference-between-brightfield-and-fluorescence-microscope/

10. When toxicity of plastic particles comes from their fluorescent dye: a preliminary study involving neotropical Physalaemus cuvieri tadpoles and polyethylene microplastics. Journal of Hazardous Materials Advances. 2022;6: 100054. doi:10.1016/j.hazadv.2022.100054

11. Cell Imaging & Analysis. Available: https://www.moleculardevices.com/applications/cell-imaging

12. Smailović J. How AI and Automation are Revolutionising Microscopy. Endava; 7 Dec 2022 [cited 17 May 2023]. Available: https://www.endava.com/en/blog/Business/2022/how-ai-and-automation-are-revolutionising-microscopy

13. von Chamier L, Laine RF, Henriques R. Artificial intelligence for microscopy: what you should know. Biochem Soc Trans. 2019;47: 1029–1040. doi:10.1042/BST20180391

14. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521: 436–444. doi:10.1038/nature14539

15. Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, et al. Deep Learning: A Primer for Radiologists. Radiographics. 2017;37: 2113–2131. doi:10.1148/rg.2017170077

16. The Basic Concepts of Machine Learning. [cited 17 May 2023]. Available: https://www.domo.com/glossary/what-are-machine-learning-basics

17. Mohri M, Rostamizadeh A, Talwalkar A. Foundations of Machine Learning, second edition. MIT Press; 2018. Available: https://play.google.com/store/books/details?id=dWB9DwAAQBAJ

18. What is Machine Learning? [cited 17 May 2023]. Available: https://www.ibm.com/topics/machine-learning

19. What is Supervised Learning? [cited 17 May 2023]. Available: https://www.ibm.com/topics/supervised-learning

20. Johnson D. Unsupervised Machine Learning: Algorithms, Types with Example. In: Guru99 [Internet]. 5 Feb 2020 [cited 17 May 2023]. Available: https://www.guru99.com/unsupervised-machine-learning.html

21. Sep 27. 9 Real-Life Examples of Reinforcement Learning. In: SCU [Internet]. Leavey School of Business, Santa Clara University; [cited 17 May 2023]. Available: https://onlinedegrees.scu.edu/media/blog/9-examples-of-reinforcement-learning

22. Brownlee J. A Gentle Introduction to Ensemble Learning Algorithms. In: MachineLearningMastery.com [Internet]. Machine Learning Mastery; 18 Apr 2021 [cited 17 May 2023]. Available: https://machinelearningmastery.com/tour-of-ensemble-learning-algorithms/

23. Reyes K. What is Deep Learning and How Does It Works [Explained]. In: Simplilearn.com [Internet]. Simplilearn; 22 Apr 2020 [cited 17 May 2023]. Available: https://www.simplilearn.com/tutorials/deep-learning-tutorial/what-is-deep-learning

24. Narasimman P. Top 26 Applications of Deep Learning in 2023. [cited 17 May 2023]. Available: https://www.knowledgehut.com/blog/data-science/deep-learning-applications

25. Yan J. Using pre-trained Vision Transformer model and ResNet model as features extractors for image popularity prediction. In: Medium [Internet]. 4 Jan 2022 [cited 17 May 2023]. Available: https://medium.com/@james.sc.yan/using-pre-trained-vision-transformer-model-and-resnet-model-as-features-extractors-for-image-2292096e99a

26. Hamdaoui Y. Image Data Analysis Using Python: In: Towards Data Science [Internet]. 8 Dec 2019 [cited 17 May 2023]. Available: https://towardsdatascience.com/image-data-analysis-using-python-edddfdf128f4

27. Cell Painting. Available: https://www.moleculardevices.com/applications/cell-imaging/cell-painting

28. PhenoVue Hoechst 33342 Nuclear Stain 100mg. [cited 17 May 2023]. Available: https://www.perkinelmer.com/product/phenovue-hoechst-33342-100mg-cp72

29. Perkin Elmer LLC PhenoVue 512 Nucleic Acid Stain. [cited 17 May 2023]. Available: https://www.fishersci.com/shop/products/phenovue-512-nucleic-acid-stai/502093531

30. PhenoVue 641 Mitochondrial Stain. [cited 17 May 2023]. Available: https://www.perkinelmer.com/product/phenovue-641-mitochondrial-stain-cp3d1

31. PhenoVue Fluor 488 - Concanavalin A. [cited 17 May 2023]. Available: https://www.perkinelmer.com/product/phenovue-fluor-488-concanavalin-a-cp94881

32. PhenoVue Fluor 555 - WGA. [cited 17 May 2023]. Available: https://www.perkinelmer.com/product/phenovue-fluor-555-wga-cp15551

33. Wikipedia contributors. Normalization (image processing). In: Wikipedia, The Free Encyclopedia [Internet]. 1 Apr 2022. Available: https://en.wikipedia.org/w/index.php?title=Normalization_(image_processing)&oldid=1080487172

34. Zach. Z-score normalization: Definition & examples. In: Statology [Internet]. 12 Aug 2021 [cited 23 May 2023]. Available: https://www.statology.org/z-score-normalization/

35. Terra J. Pytorch vs Tensorflow vs Keras: Here are the difference you should know. In: Simplilearn.com [Internet]. Simplilearn; 27 Jul 2020 [cited 20 May 2023]. Available: https://www.simplilearn.com/keras-vs-tensorflow-vs-pytorch-article

36. NVIDIA CUDA deep neural network (cuDNN). In: NVIDIA Developer [Internet]. 2 Sep 2014 [cited 20 May 2023]. Available: https://developer.nvidia.com/cudnn

37. GPUs for machine learning. In: IT Connect [Internet]. [cited 20 May 2023]. Available: https://itconnect.uw.edu/guides-by-topic/research/research-computing/gpus-for-machine-learning/

38. HPC Center. HPC Center. In: HPC Center [Internet]. [cited 20 May 2023]. Available: https://hpc.ut.ee

39. HPC Center. Rocket. In: HPC Center [Internet]. [cited 20 May 2023]. Available: https://hpc.ut.ee/services/HPC-services/Rocket

40. Project jupyter. [cited 20 May 2023]. Available: https://jupyter.org/

41. Image Classification Using ResNet-50 Network — MindSpore r1.1 documentation. [cited 17 May 2023]. Available: https://www.mindspore.cn/tutorial/training/en/r1.1/advanced_use/cv_resnet50.html

42. Bhojanapalli S, Chakrabarti A, Glasner D, Veit A, Song C, Wu J, et al. Figure 1. Transformers vs. ResNets. While they achieve similar. In: ResearchGate [Internet]. [cited 17 May 2023]. Available: https://www.researchgate.net/figure/Transformers-vs-ResNets-While-they-achieve-similar-performance-for-image_fig1_350457424

43. Danielsen N. Simple Image Classification with ResNet-50 - Nina Danielsen. In: Medium [Internet]. 22 Nov 2019 [cited 17 May 2023]. Available: https://medium.com/@nina95dan/simple-image-classification-with-resnet-50-334366e7311a

44. Website. Available: https://en.wikipedia.org/wiki/Residual_neural_network

45. Ramaswamy S. What is the Vision Transformer Model (ViT) and How Does it Help in Image Recognition? In: Akaike AI [Internet]. Akaike Tech; 27 Mar 2023 [cited 23 May 2023]. Available: https://akaike.ai/what-is-the-vision-transformer-model-vit/

46. Brownlee J. Autoencoder feature extraction for classification. Machine Learning Mastery December. 2020;6: 2020. Available: https://machinelearningmastery.com/autoencoder-for-classification/

47. Moparthi S. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (Vision Transformers). In: Analytics Vidhya [Internet]. 10 Mar 2021 [cited 23 May 2023]. Available: https://www.analyticsvidhya.com/blog/2021/03/an-image-is-worth-16x16-words-transformers-for-image-recognition-at-scale-vision-transformers/

48. google/vit-large-patch16-224-in21k · Hugging Face. [cited 23 May 2023]. Available: https://huggingface.co/google/vit-large-patch16-224-in21k

49. Swin Transformer for hierarchical vision (2021) - KiKaBeN. In: KiKaBeN - Smart Tech Information: From Concept to Coding [Internet]. KiKaBeN; 4 Nov 2022 [cited 23 May 2023]. Available: https://kikaben.com/swin-transformer-2021/

50. microsoft/swin-large-patch4-window7-224 · Hugging Face. [cited 24 May 2023]. Available: https://huggingface.co/microsoft/swin-large-patch4-window7-224

51. Berryman S, Matthews K, Lee JH, Duffy SP, Ma H. Image-based phenotyping of disaggregated cells using deep learning. Commun Biol. 2020;3: 674. doi:10.1038/s42003-020-01399-x

NON-EXCLUSIVE LICENCE TO REPRODUCE THESIS AND MAKE THESIS PUBLIC

I, *Ali Zeynalli*,

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,

   Computer Vision Meets Microbiology: Deep Learning Algorithms for Classifying Cell Treatments in Microscopy Images

   supervised by Asst. Prof., PhD Dmytro Fishman

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.

Ali Zeynalli

24/05/2023