# Utilizing convolutional neural networks for data-driven modelling of stochastic processes with application to stem cell differentiation

Josh Chang

Department of Mechanical Engineering

McGill University, Montreal

August 2024

Supervised by Dr. Michael Kokkolaras

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

# Acknowledgements

Firstly, I would like to thank my graduate supervisor, Dr. Michael Kokkolaras, for his supervision over the last two years and his critical support which allowed for this thesis to be completed. I came to McGill to pursue a graduate degree with you to broaden my capability to solve engineering problems and I have certainly enjoyed the time spent here.

Secondly, I am grateful for the support from the Stem Cell Bioprocessing Laboratory led by Dr. Corinne Hoesli and particularly Dr. Hamid Ebrahimi Orimi and Dr. Jonathan Brassard. You were instrumental in providing data and guidance on how to look at and interpret stem cells and helped me tackle a problem in a field that was very new to me.

I would also like to thank Dr. Jeremy Laliberte and Dr. Iryna Borshchova who have given me research advice and supervision primarily during my undergraduate research work. Both of you had a profound and significant impact on me through your guidance and encouraging me to do a postgraduate degree. Additionally, I appreciated the thorough and insightful questions by Dr. Khalil Al Handawi and the advice you have given me through my master's. I also greatly appreciate the conversations made with all members of the Systems Optimization Lab regarding our respective research topics.

I would also like to acknowledge the financial support of Médicament Québec for this research topic. Additionally, I would like to thank the contributions of the Natural Sciences and Engineering Research Council of Canada through the Canada Graduate Scholarships – Master's (CGS-M) scholarship for my work on optimization in aerospace.

Finally, I would like to thank my parents, family, and friends for their stalwart support of me and these efforts throughout the last two years.

# Abstract

The creation of pancreatic islets from stem cells is a multi-stage biological methodology, referred to as a bioprocess, which has high variability and limited explainability. A successful bioprocess must have a sufficiently high flow cytometry score and the high cost and destructive nature of this test makes it infeasible for production at scale. Therefore, reducing the cost and wastage can be achieved through modelling techniques which can approximate or estimate the flow cytometry score and allow for quality control to occur. In this thesis, the application of analytical and opaque surrogate models is trained on stem cell images to predict potentially relevant visual phenomena, referred to as biomarkers, and flow cytometry itself. Transparent models are shown to have insufficient capacity to properly approximate biomarkers with proposed biomarkers in the literature being insufficient to estimate flow cytometry. However, a deep convolutional neural network is successfully able to make predictions on flow cytometry to suggest avenues for future research.

# Résumé

La création d'îlots pancréatiques à partir de cellules souches est une méthodologie biologique en plusieurs étapes, appelée bioprocédé, qui présente une grande variabilité et une explicabilité limitée. Un bioprocédé réussi doit avoir un score de cytométrie de flux suffisamment élevé et le coût élevé et la nature destructrice de ce test le rendent impossible à produire à l'échelle. Par conséquent, la réduction des coûts et du gaspillage peut être obtenue grâce à des techniques de modélisation qui peuvent approximer ou estimer le score de cytométrie de flux et permettre un contrôle de la qualité. Dans cette thèse, l'application de modèles de substitution analytiques et opaques est entraînée sur des images de cellules souches pour prédire des phénomènes visuels potentiellement pertinents, appelés biomarqueurs, et la cytométrie de flux elle-même. Les modèles transparents s'avèrent insuffisants pour approximer correctement les biomarqueurs, les biomarqueurs proposés dans la littérature étant insuffisants pour estimer la cytométrie de flux. Cependant, un réseau neuronal convolutionnel profond est capable de faire des prédictions sur la cytométrie de flux et de suggérer des pistes pour la recherche future.

# Contribution of Author

This thesis represents the sole work of the primary author supervised by Dr. Michael Kokkolaras. Experiments were done externally by the Stem Cell Bioprocessing Laboratory at McGill University. Discussions on biomarkers to consider and the existing bioprocess were done in collaboration with Dr. Hamid Ebrahimi Orimi and Dr. Jonathan Brassard.

# Contents

# List of Figures

# List of Tables

# Nomenclature and Acronyms

| | |
|---|---|
| $\mathbb{E}$ | Expectation |
| $\Theta$ | Unknown distribution |
| $I$ | Image |
| $c_d$ | Cell density |
| $c_{sa}$ | Cell surface area |
| $z$ | Action |
| ADAM | Adaptive Moment Optimizer |
| BBO | Black box optimization |
| CNN | Convolutional neural network |
| DOE | Design of experiments |
| GA | Genetic algorithim |
| ML | Machine learning |
| NOMAD | Nonlinear Optimization by Mesh Adaptive Direct Search |
| PID | Proportional-integral-derivative control |
| R-CNN | Region-based convolutional neural network |
| SCBL | Stem Cell Bioprocessing Laboratory, McGill University |
| SNN | Stacked neural network |
| YOLO | You Only Look Once |

# Chapter 1

# Introduction

In medical science, attempts to explain human disease and the potential effectiveness of treatments benefit from accurate models. As human testing is limited in number and mired with ethical considerations, stand-in systems are critical [1]. Therefore, researchers have created methodologies, referred to as bioprocesses, to create surrogate models known as organoids which replicate human organ behaviour. Organoids can be created from human pluripotent stem cells (hPSC) and act as small functional constituent parts of a full-sized organ [2] [3]. The creation of these organoids involves cell specialization, or a differentiation protocol, where hPSCs undergo a multi-stage process to become a particular cell type or system. In the thesis, we consider specifically the organoids of the pancreas and its differentiation protocols.

Pancreatic organoid protocols typically focus on the differentiation of hPSCs to functional islets which contain alpha-cells and beta-cells which collectively regulate glucose in the body [2]. Alpha-cells release glucagon to increase blood glucose while beta-cells counteract this effect through insulin production. Applications of these islets have been considered for modelling ductal pancreatic cancer [4] and applications for clinical transplantation for diabetes [5]. However, existing bioprocesses are limited in quantity and availability as the usage of

islets from multiple cadavers is required to treat a single patient. If done at a high enough consistency and quality, these bioprocesses could provide significant biological material for testing and clinical applications for human transplantation [2, 3, 5].

Despite these applications, issues remain with existing differentiation protocols with the production of organoid islets occurring in limited quantities, having limited control over the composition of the cell population that is produced [6], and the limited number of created cells being inconsistent in quantity [2]. These issues could be resolved with a robust pancreatic differentiation protocol which produces sufficient islets for wider usage. Therefore, proper modelling and optimization of the bioprocess to create these organoids could allow for formal optimization to occur, allowing for the quantitative assessment of these biomarkers.

## 1.1 Pancreatic Differentiation

Pancreatic differentiation protocols generally consist of a seeding stage followed by several numbered stages. In the initial seeding stage, a variable number of cells are added by the experimenter into the Petri dish or a cell density $c_d$ value. Figure 1.1 shows an image under a microscope at 4x magnification of the seeded hPSC after 24 hours of seeding.

In this image, two distinct textures are noted. Spaces which are occupied by the hPSC have a dotted or circular texture while empty spaces, examples of which are circled in red, appear as smooth patches. These empty spaces in the Petri dish appear as grey patches and white lines at 24 hours and 48 hours of seeding respectively. In the numbered stages, growth factors are added at each stage and can be varied temporally and in the concentration in which they are added. The time in which growth factors are added is adjusted both by when they are added at each stage and also by the quantity of time they are utilized. In each of

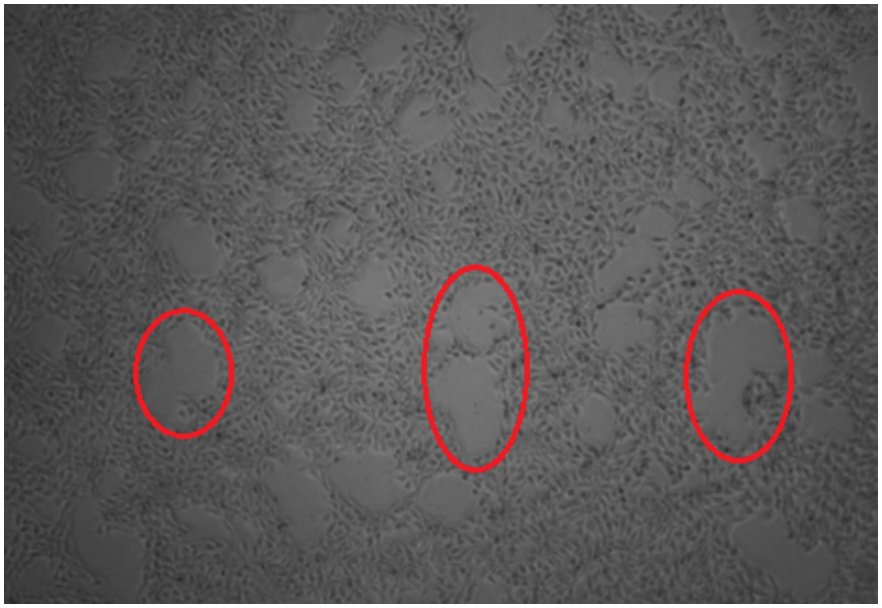


Figure 1.1: Stem cell seeding 24 hours after seeding at 4x magnification.

these stages, the total proportion of cells which successfully enter the next stage is defined as the success rate and also referred to as the differentiation rate. After all stages, the target output for the considered bioprocess is a mature beta cell which produces or exhibits insulin.

The modelling of pancreatic differentiation protocols has primarily been done using a biological approach. This includes an arbitrary number of stages between the starting stem cell and the final beta cell. At each stage, the "input" is treated as the "output" of the prior stage, which is the cell type which has been differentiated up to that point. While a 7 stage process is used both by Hoesli and collaborators and in Petersen et al. [7] the definitions of the cell types after each stage vary. The number of stages can also vary as observed in Augsornworawat et al. [5] and Sharon et al. [6] which use a 6-step process with the final stage representing beta cells. This lack of standardization results in differences in research results as noted in Casamitjana et al. [2] for very similar methodologies.

Despite the methodological differences in defining stages and inputs, output definitions

are highly consistent across research groups. Single-cell sequencing is used to identify the number of cells which exhibit a certain gene expression, or behaviour, which are common to the target cell type. An example of this would be the percentage of cells which exhibit PDX1 which has been explored extensively in research [8–12] and used to define the "cellular identity" of the outputs at each stage. This rate can be defined as a "success rate" for the experiment to maximize this value. This success rate can be measured using flow cytometry. Flow cytometry measures the presence and quantity of specific biomarkers which are known to be exhibited or not exhibited by certain cell types.

Despite the unified approach to quantifying differentiation success rates, very low success rates are currently observed in research approaches. At the final stage of each protocol, which involves the generation of mature beta cells, success rates vary from 30% in Sharon et al. [6] and between 20-35% in the Stem Cell Bioprocessing Laboratory (SCBL) at McGill. Therefore, an improvement in this success rate could benefit research and clinical applications which depend upon differentiated beta cells. In this thesis, the modelling techniques to quantify biomarkers and causes of cell differentiation failure are explored. This is achieved by considering the differentiation protocol utilized in the SCBL and creating models to approximate and perform an a priori prediction of the flow cytometry value. A summary of the pancreatic differentiation process considered is shown in Figure 1.2.

The inputs and outputs are shown in Figure 1.2 from biological and mathematical modelling perspectives. From a biological perspective, this thesis assumes stem cells are allowed to expand into confluent stem cells followed by the 7 stages of differentiation. In contrast, the modelling approach typically involves the measurement of cell density or the number of cells per unit area, the cell line or origin of the stem cells, and could include the cell surface area which defines the total occupied space of the stem cells.

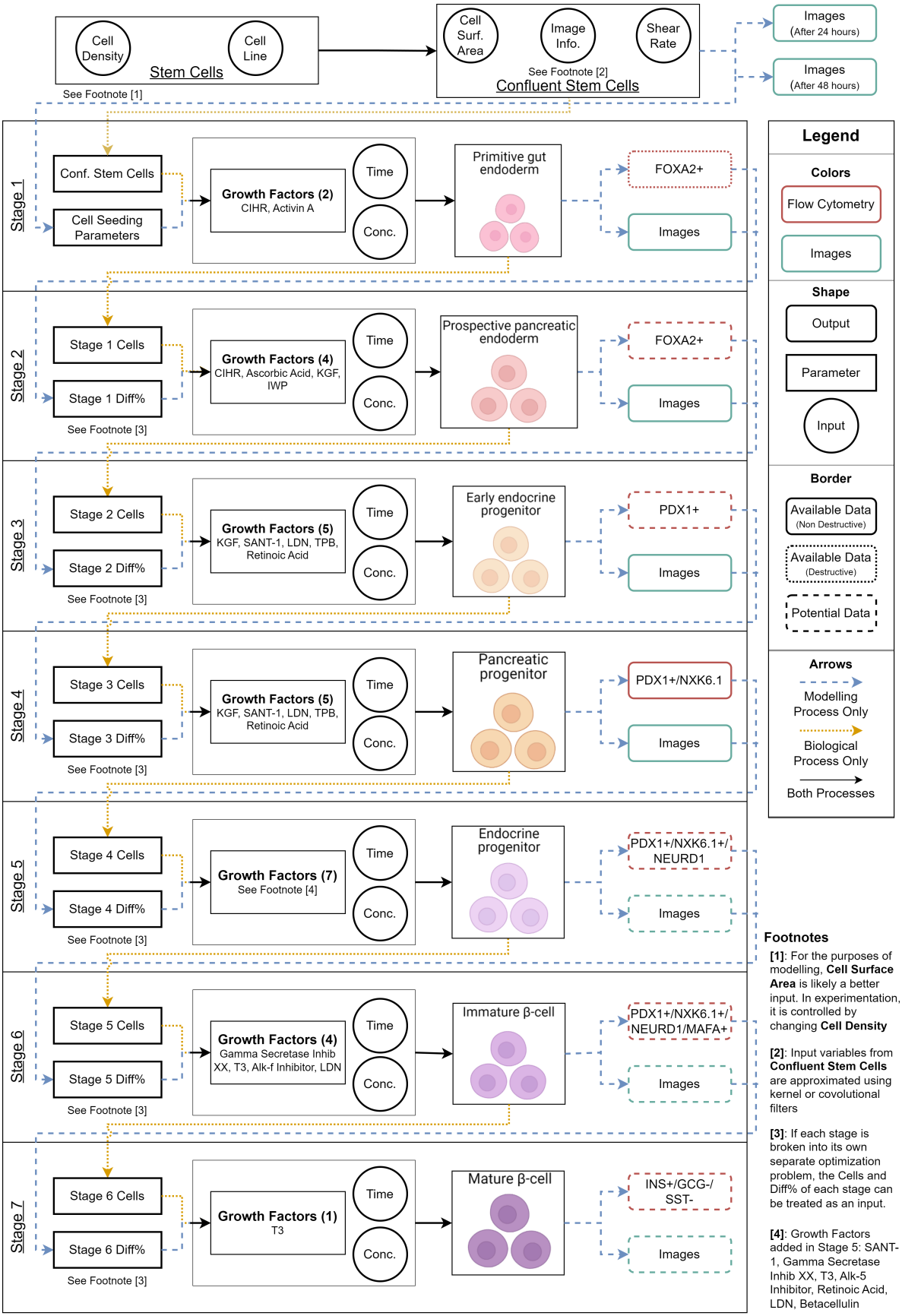



Figure 1.2: Block diagram of the bioprocess from modelling and biological views.

To improve this protocol, a trivial solution may be to run a sufficient number of experiments to model the potential impacts of each variance in input. However, the biological differentiation protocol suffers from certain difficulties which can be broadly classified into three categories: experimental quantity, biological modelling, and robustness.

### 1.1.1 Experimental Quantity

For the pancreatic bioprocesses, a significant limitation is present in the number of available experiments which correlate directly to the number of "function evaluations" for a given algorithm. Biological experiments are slow and require human intervention throughout which limits the number of sequential experiments. In Figure 1.2, the seeding stages take up to 2 days, followed by 14 days to finish Stage 4. Financial limitations due to the high cost of growth factors limit the number of experiments which are done in parallel by experimenters. However, it is noted that recent research includes the exploration of certain additives to substitute expensive growth factors [13]. However, with the variance in bioprocessing methodologies, the applicability of these substitutions to other research groups may be of limited use.

While automated tools and machines have been developed to assist humans in scaling these bioprocesses, such as the Ambr 15 machine [14], significant human intervention is still required. These difficulties have been observed by the author in practice. In just over a year of research, the author notes that under 30 wells have been done providing a significant limitation on the approaches which can be taken.

### 1.1.2 Biological Modelling

The pancreatic bioprocess is inconsistently modelled throughout the existing literature, although most protocols have similarities in the way they are modelled. The potential inputs of

a bioprocess are high in number and vary in numerical type as shown in Figure 1.2. Outputs are generally a single number or distribution of cells which is modelled as a single objective with no trade-off.

For the seeding stage, a decision must be made on the type of cells used, the number of cells used, and the quantity of time provided for seeding. Additionally, in each of the 7 stages, there are between 1 and 7 growth factors which are applied for different amounts of time, concentrations, and at a different period. While the type of cells and type of growth factors used can be treated as categorical variables or variables which can take one of a few discrete options, the remainder are continuous variables. The large number of mixed variable types limits the types of optimization approaches which can be considered.

For the output, flow cytometry is capable of providing continuous numerical outputs, albeit destructive when cells are not in suspension, or floating in liquid media. When performed, the percentage of successful gene expressions or differentiation rate can be measured which provides a normalized value between 0 to 100% of the number of cells which have successfully specialized. In contrast, all single-cell sequencing tests are destructive but provide highly detailed information on the attributes of the cell behaviour. Therefore, a significant observer effect is present with the quantification of experimental quality resulting in the loss of potentially successful differentiations. As the tested cells are destroyed, in general, multiple experiments are done in parallel with the assumption all experiments have approximately equal quality.

The destructive nature of this test presents two issues. Firstly, the assumption that all experiments on the same plate have approximately equal quality means there is no consideration for plate-specific phenomena. The fluid shear stress applied by experimenters regulates

pancreatic development [15] and varies from well to well but cannot be modelled if these wells are assumed to be similar and only modelled with biological inputs. Secondly, the destruction of wells to model prior stages means a single experiment can only assist in a single modelled stage. As cells are primarily useful in the later stages when they reach maturity, these optimization loops would have to occur for each stage to allow the greatest percentage of cells to survive till later stages.

### 1.1.3 Robustness

Differentiation protocols currently lack robustness and definable parameters to determine quantitative feasibility as noted in [2, 3]. Methodologies generally have limited reproducibility, have highly variable success rates, and potential indicators for failure are not currently considered or observed. Therefore, similar runs of the same protocol are stochastic with potentially significant variance with the same set of input variables. While the primary objective of bioprocess optimization is generally to improve the mean value of the methodology, an eventual capability to reduce the variance is likely beneficial.

## 1.2 Difficulties in Optimization

These three highlighted attributes of the bioprocess modelling and optimization problem introduce numerical difficulties when attempting to optimize. These difficulties either increase the number of required experiments or decrease the number of available experiments. For modelling, the presence of discrete variables introduces discontinuities which make simple surrogate modelling techniques difficult. Response surface methodologies with a design of experiments can allow for the building of a quadratic model and permit optimization using derivative-based methods to find local minima or potentially optimal solutions [16].

When engineering systems have certain mathematical properties, such as smoothness, existing gradient-based optimization methodologies can be applied to find a locally optimal set of inputs. However, discrete variables and functions introduce discontinuities which require problem relaxation, derivative-free, or black-box optimization.

The optimization of any complex processes requires the usage of 'function evaluations' where the output of a system of interest given input is considered. Given the limited number of function evaluations, or flow cytometry tests, forming a biological model coupled from the biological inputs to the differentiation percentage is likely infeasible. State-of-the-art black-box optimization algorithms, such as NOMAD 4 [17], require large numbers of function evaluations. To optimize a single stage of 1-5 variables, the order of magnitude would be approximately a hundred to thousand times the total variable count. This is well outside the feasible number of experiments which could be realistically performed as mentioned in discussed in Section 1.1.1

## 1.3 Motivation

While significant difficulties exist in modelling an opaque process with limited data, many downsides can be reduced or minimized with the creation of approximating models. Financial cost and data collection time, which can take weeks, can both be reduced. Additionally, they can aid in the decision-making of the experimenter, which forms a baseline to compare against.

These models can be data-driven, and rely upon historical experiments, or "physics-based" and rely upon simulation. However, physics-based modelling for this problem is

infeasible due to the lack of an underlying understanding of the stem cell differentiation process. Existing approaches to creating digital twins or virtual simulators are often highly rudimentary with a severe lack of robustness to either quantify the success of a hypothetical set of inputs or quantify the likelihood of success for a specific experiment.

Therefore, at present, with imperfect knowledge of the process, a data-driven model to assist present-day bioprocesses is beneficial. This includes exploring both transparent models and black-box models up to and including deep learning. The data processing pipeline and models are designed with the variability of the bioprocess in mind and the understanding that limited data is available.

# Chapter 2

# Background

The literature review is separated into two subsections. Firstly, the methodologies used to improve pancreatic differentiations are discussed in Section 2.1. Next, Section 2.2 reviews the models and optimization frameworks for process control and process optimization in mechanical engineering. Particular emphasis is placed upon manufacturing processes which are infrequent and expensive and mirror the problems highlighted in Section 1.2.

## 2.1 Pancreatic Differentiation Improvements

Pancreatic differentiation improvements have been done commonly through informal "optimization" methods. Two different approaches to improving the differentiation rate are discussed: Firstly, variations to cell density and aggregation are described in Section 2.1.1 and secondly, variations of the growth factors used in differentiation are discussed in Section 2.1.2.

It is noted that these existing approaches involve a priori improvements where changes to the input variables are proposed based on prior biological knowledge or intuition. A comparison of average differentiation quality is performed between two different sets of inputs

followed by a statistical test using its mean and variance. While statistical testing can prove with particular confidence the new set of inputs produces an improved result, an ability to quantify the stochastic noise with non-biological inputs may reduce the variance and be more relevant to individual samples.

### 2.1.1 Cell Density

Attempts to improve the rate of differentiation have included variations of the available variables shown in Figure 1.2. Gage et al. [18] identify cell cultures seeded at high cell densities that increased the number of cells which exhibit insulin and glucagon, which are indicators of mature beta cells or the target cell output. Similar findings are observed in Takizawa-Shirasawa et al. [19] where the combination of fibroblast growth factor 7 (FGF7) and high cell density improved the differentiation success. Toyoda et al. [20] hypothesize the improvements in differentiation success were a result of the greater signalling capabilities of aggregated cells. Improvements based upon this theory are observed in Tran et al. [11] where micropatterns are applied to the well surface. These patterns improve local clustering and aggregation of cells in the orientation in which the patterns are placed.

For these reasons, the period of cell growth, or confluence, where cells expand in surface area, the well-empty area in the Petri dish is believed to be highly significant to the success of the differentiation process. However, it is plausible the quantity of surface area covered could be more correlated than cell density. It is believed the same experimental value of $c_d$ can result in well-empty area values with a 10-20% difference even when the same amount of time is cell seeding. The well-empty area, or area not occupied by the cell surface area $c_{sa}$, is the primary visual cue currently used to define differentiation quality for human experimenters in the SCBL at McGill University. Anecdotally, this approach is used to improve the differentiation rate by identifying when cells are ready for the following differentiation steps.

The author is not aware of research which defines the maximum number of cells permissible before decreases in the differentiation rate occur.

### 2.1.2 Growth Factors, Additives and Markers

Growth factors added in at each stage vary in type, concentration, and temporal period. These factors are intended to inhibit or exhibit certain signals at each stage of the bioprocess to maximize the differentiation rate. In Stage 1 of Figure 1.2, where differentiation aims to create endoderm cells, D'Amour et al. [21] identifies the necessity of Acitvin A. Similar findings by Xu et al. [22] propose a 3-4 day application period of Activin A and two additional growth factors to create cells with significant PDX1 expression. Following this, Ghorbani-Dalini et al. [23] quantifies an improved Activin A concentration in addition to a new knockout serum (KSR) which replaces existing additives to improve differentiation efficiency. However, the significant applications of Activin A in the early stages result in significant financial costs [13]. Therefore, Jiang et al. [13] proposes a modified protocol which uses small molecules instead of Activin A and greatly reduces the cost of this individual step. Despite its applications in the first stage of cell differentiation, Cho et al. [24] find the suppression of activin to be beneficial in the differentiation from the endoderm cells to the pancreatic endoderm in Stage 2 shown in Figure 1.2.

However, the significant applications of Activin A in the early stages result in significant financial costs [13]. In this later stage, Retinoic acid is found to induce PDX1 exhibition [24, 25] with the signalling and mechanisms in which this occurs explained in Loberbaum et al. [26]. In these papers, the Ribonucleic acid (RNA) of a cell is identified using single-cell sequencing and shows the different stages of cell development. These methodologies have been used to identify the relationship between PDX1 and cell differentiation [8–12], that

various paths can be taken to achieve the same differentiated cells [7, 27], and alternative ways to mature cells [5].

## 2.2 Process and Quality Control

Ways to address these manufacturing issues have been explored in bioprocess process control and quality control. Research on this topic has been summarized in the review paper Alford [28]. This paper describes the impact digitization has played in modern biological process control through the monitoring and management of biological inputs such as "pH, temperature, and dissolved oxygen". While biological research has proposed various proportional-integral-derivative (PID) control systems to tackle manufacturing plants, these assume a stable methodology which is monitored and adjusted in the loop to ensure a stable response. Rathore et al. [29] describe these PID control schemes which can tackle variability and nonlinearity. However, no feedback can be provided to the system as flow cytometry testing is typically destructive.

Therefore, a stable methodology or a surrogate feedback model which provides information to a process control system must be designed. For quality control and improvements, attempts to optimize this process have included the design of experiments (DOE) and genetic algorithms (GA). However, these approaches require exponentially increasing experiments when the variable number increases. Similar approaches described in Mondal et al. [30] involve machine learning (ML) based approaches which include supervised approaches, where labelled data is required, and unsupervised approaches where relations are inferred. However, the mechanisms which can be modelled either require data quantities beyond which the SCBL can afford for pancreatic stem cells or lack the expressiveness to model the behaviours.

To tackle this problem, inspiration is taken from mechanical engineering problems where manufacturing processes are involved. Process control and statistical modelling are seen frequently on the manufacturing lines of automotive and aircraft components. In the author's opinion, aerospace engineering manufacturing has similar issues to bioprocess optimization with low volumes, high costs, and long times for each development process. Similarly to the Rathore et al. [29] systematic review, Li et al. [31] summarize 9 years of proposed approaches in "smart manufacturing" primarily in an engineering context and highlight the effectiveness of reinforcement learning algorithms to make online adjustments.

Similar approaches have been adopted in additive manufacturing as described by Kumar et al. [32] leading to "industry 4.0" where data is collected and interlinked. Approaches to tackling the high cost of experiments have been tackled through the idea of "digital twin" models where a simulated system emulates real-life experiments [33]. While the pancreatic bioprocess is not amenable to simulation-based approaches, such as finite element analysis (FEA) in manufacturing, we can similarly utilize prior knowledge to form estimations of reality. This can be achieved through collected images as described in Figure 1.2 which could theoretically be taken continuously and used to provide a stable feedback loop for quality control.

## 2.3 Image Analysis and Feature Detection

In this thesis, the identification of particular features in images is discussed at three separate complexity levels. Firstly, in Section 2.3.1 white-box or transparent approaches as described. Secondly, black-box machine learning (ML) approaches are explained in Section

2.3.2. Finally, Section 2.3.3 describes object detection ML approaches.

### 2.3.1 White Box or Transparent Approaches

Existing biological approaches to image processing and feature localization include ImageJ [34] and CellProfiler [35]. These software have built-in quantitative methods to assess the number of cells and texture in addition to preprocessing techniques specific to microscope images. This can include common issues with digital microscopes such as "vignetting" where lighting varies between the center and edges of an image [36]. When combined these tools have been used to build "pipelines" which have been used to count tumours and different cell types [35].

Common tools in mechanical engineering include the MATrix LABoratory (MATLAB) which contains the Image Processing Toolbox [37]. This library allows for the preprocessing, analysis, and segmentation of various types of data. Approaches to achieve these tasks include edge detection, which can be done using a convolution across the image using a kernel. This approach can be used to model and estimate these three-dimensional surfaces as described in Barrow and Tenenbaum [38].

This kernel can be shaped with an approach which involves morphological "structuring elements" or "strel" [39] to perform background removal to solve the vignetting problem [36, 37]. These shapes can also be used to standardize images and assist with imperfections such as out-of-focus images through sharpening and increases in contrast [40].

### 2.3.2 Black Box Machine Learning

ML encompasses a wide variety of white and black box applied statistics which can be used to identify a target variable or supervised tasks. Convolutional Neural Networks (CNNs) are built upon the assumption that 1) the order of inputs affects the output, 2) local phenomena

in an image are more relevant than far-range phenomena, and 3) an invariance in feature location or where something is observed [41]. In general, the assumptions made in these networks are that convolutional layers are intended for "feature extraction" or identification of relevant features. This is then followed by multiple fully connected layers which are intended to model the target function based upon these features.

### 2.3.3 Object Detection

Object detection, also referred to as image segmentation, describes a variety of approaches where the objective is to identify the location and size of specific phenomena. These methodologies are typically achieved through black-box ML approaches and generally have a trade-off between inference speed and mean average precision (mAP). They are designed for portable or online processes where an immediate determination of the observed phenomena is necessary and would therefore be suitable for a manufacturing process where immediate feedback is required. These approaches have been used for obstacle avoidance or detect and avoid algorithms in drones.

Implementation of these black box approaches for object detection are very numerous with only a few described in this summary in ascending order of complexity. The previously mentioned Image Processing Toolbox from MATLAB [37] contains the active contour tool which attempts to segment objects from a background. Newer algorithms such as mask region-based convolutional neural networks (Mask R-CNN) improve upon this resolution as they perform image segmentation and provide "masks". These "masks" include coverage maps of specific phenomena instead of a simple box outline [42]. Finally, the You Only Look Once (YOLO) algorithm forms the foundation of numerous subsets of these algorithms with a novel approach to scanning the image to reduce inference times. The geometric center of the phenomena, referred to as anchor points, and the size of the object in a box form

are provided as outputs to a trained YOLO algorithm [43]. The author notes that many variants of the YOLO models have been designed which can vary dramatically from this very simplified description.

## 2.4 Motivation and Proposed Approach

While numerous research using single-cell profiling have attempted to explain the steps and pathways in which certain cells grow, the author is unaware of research attempting to quantify an experiment's feasibility through historical data. Additionally, in contrast to the approach taken by existing biological researchers, the principle of "process control" seen in mechanical engineering is proposed. Therefore, in this thesis, a methodology which monitors and quantifies the state of a given differentiation is proposed.

This quantification could be used to: 1) define if a differentiation is feasible or provide a flow cytometry estimation, 2) determine if a differentiation is ready to proceed to the next stage, and 3) quantify the improvement or harm due to changes in the bioprocess methodology. To achieve this, a surrogate model is proposed over the differentiation process to predict the likelihood of success of cells at a future point in time for a given "input" using the image information.

# Chapter 3

# Methodology

In the proposed methodology, two steps are required to improve stem cell differentiation. The first step is the creation of a surrogate model with the capability to quantify experimental quality from images to differentiation percentages. The second step involves applying this model to compare the impact of various growth factors from a fixed starting point. This thesis will develop the methodology for the first step and introduce the second step as a potential application.

This thesis describes the overall problem formulation in addition to a methodology for the former step, namely, from data collection to predicting flow cytometry values. Section 3 describes the data collection approach taken by the SCBL in collaboration with the author.

## 3.1 Data Collection

The data collection methodology describes the existing data collection and the changes made to collect the necessary data. Each experiment varies a portion of the available optimization parameters in Figure 1.2. Figure 3.1 shows the components which make up an experiment.

In the current data collection methodology, flow cytometry values, $y$ are currently col-

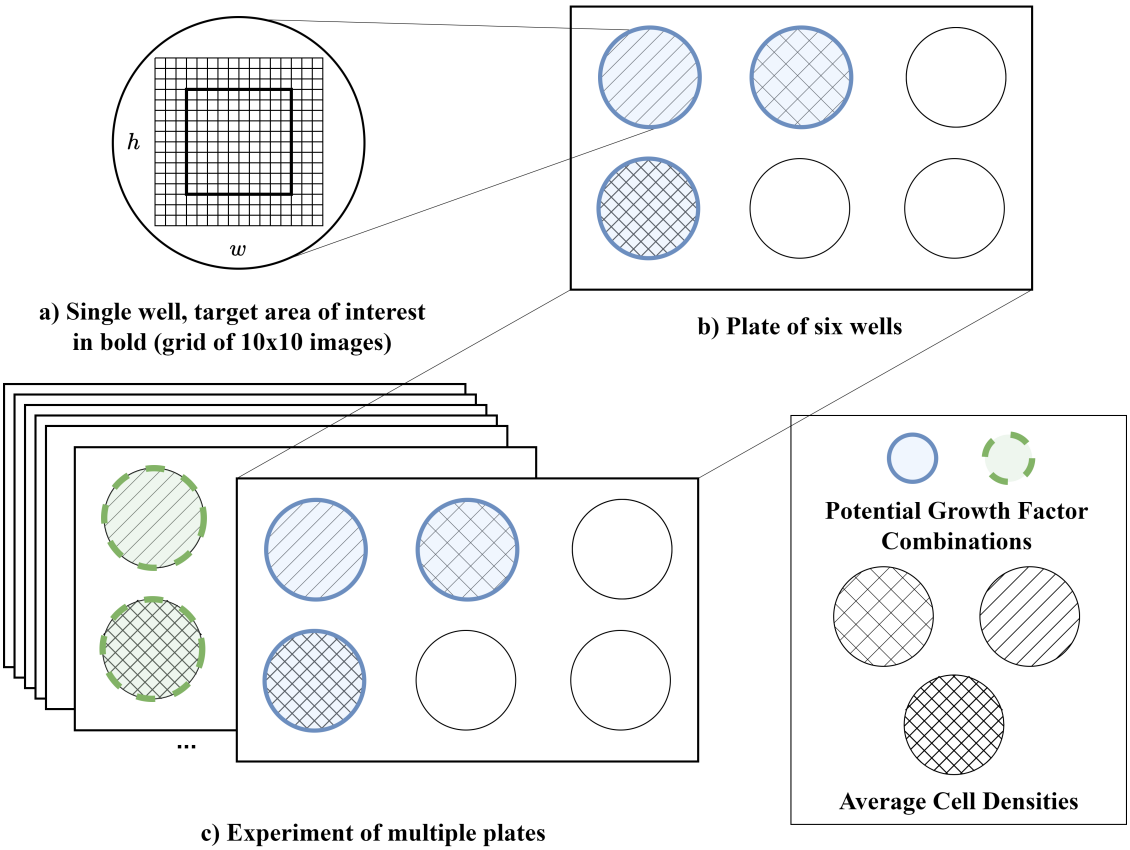



Figure 3.1: Block diagram of the bioprocess from modelling and biological views.

lected for stages 1 and 4. Stage 1 flow cytometry is a destructive test and therefore the cells are destroyed after measurement. Each well can have between 80 to 144 unique images which take upwards of 20 minutes to collect. These images are collected in a grid, which is sized arbitrarily between 8×10 and 12×12, and are taken approximately 24 hours and 48 hours after seeding. Each image which is taken has variable pictorial features including brightness, contrast, and sharpness.

For this reason, a standardization process is used similarly to an "illumination correction" in CellProfiler [35]. MATLAB is used to provide more fine-tuned control over the process and the author's existing familiarity with the tool. After reading the image, four steps are performed: Firstly, pixels above the 95th percentile are scaled to have an average intensity

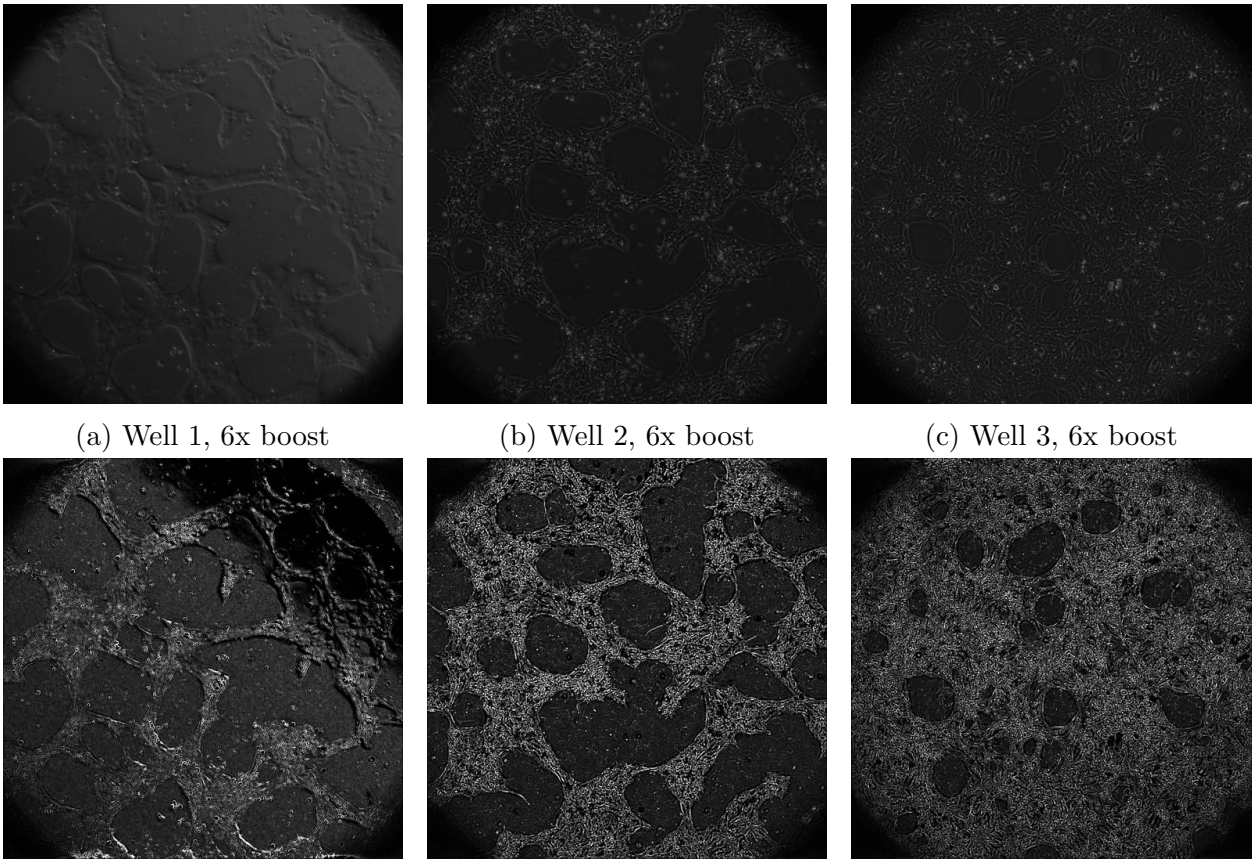
of 12% (or a value of 8,000). This first step is intended the image brightness from being proportional to well-empty area which is dark. If not adjusted, the image can appear overexposed and distinctly different to other images. Secondly, the average brightness of the image is raised by a multiplier $b$ to emphasize image-based features for the second step. Thirdly, a strel disk with radius $r$ is used with the imbothat function which performs bottom-hat filtering [37]. Finally, pixels with a brightness above a 1.5% intensity (or a value above $t_{cap}$ for an unsigned 16-bit integer) are scaled to have an average intensity of 24% (or a value of 16,000).

The first and final steps are intended to separate the lines from the empty area and standardize their brightness. The second step of increasing the overall image brightness is primarily to ensure the imbothat function has a significant enough variance between empty area and line brightness to remove the background properly.

A visual comparison of the impact of the preprocessing is shown in Figure 3.2. On the top row, Figures 3.2a, 3.2b, and 3.2c show the original image with a 6x multiplier to the pixel values. Figures 3.2d, 3.2e, and 3.2f use the preprocessing with settings of $b = 5$, $r = 15$, and $t_{cap} = 1000$

This variance is partially caused by the three factors. Firstly, limited use of the automatic focusing on the microscope is done to conserve time [1] which results in images not being in ideal focus. Secondly, the design limitations of the low tolerance composite 3D-printed bracket for the microscope reduce the image quality as the wells are at an angle to the table. This means that certain edge images, such as Figure 3.2a are significantly brighter than images closer to the center of the well such as Figures 3.2b and 3.2c. Finally, the meniscus effect of the liquid in the wells results in a difference in image quality between images taken

[1] While a greater usage of automatic focus would improve the image quality, the negative effects of leaving stem cells in this less controlled environment could result in cell death and contamination.



(a) Well 1, 6x boost (b) Well 2, 6x boost (c) Well 3, 6x boost

(d) Well 1, preprocessed (e) Well 2, preprocessed (f) Well 3, preprocessed

Figure 3.2: Comparison of raw images at 24 hours and preprocessing standardization

closer to the center versus the edge of the wells. However, the usage of the preprocessing methodology as seen in Figures 3.2d, 3.2e, and 3.2f is effective at removing the variance in lighting and emphasizes the lines and has limited effectiveness in improving issues regarding focus.

Due to the highly variable and experimental nature of the data, the problem input is effectively stochastic with similar inputs leading to variable outputs. For this reason, the number of images which may be required to define an "input" is unknown. While 80 to 144 unique images are taken per well, no existing literature defines how much well area or how many images are required to obtain a representative view of the well, experiment, or biomarker.

## 3.2 Modelling Methodology

Bioprocess modelling in this thesis is defined as an attempt to find a relation between flow cytometry values $y$ and each "set" of images $x_i$, which both have a distribution $\Theta$ which is unknown. The total number of images must be divided by the size of each "set" to define the amount of unique input and training data available to the system. This relation can be formed in three separate methods summarized in Figure 3.3.

An indirect approach involves the estimation of a visual biomarker or feature of interest to estimate flow cytometry. In contrast, a direct approach is defined as one where flow cytometry is directly estimated from the source image.

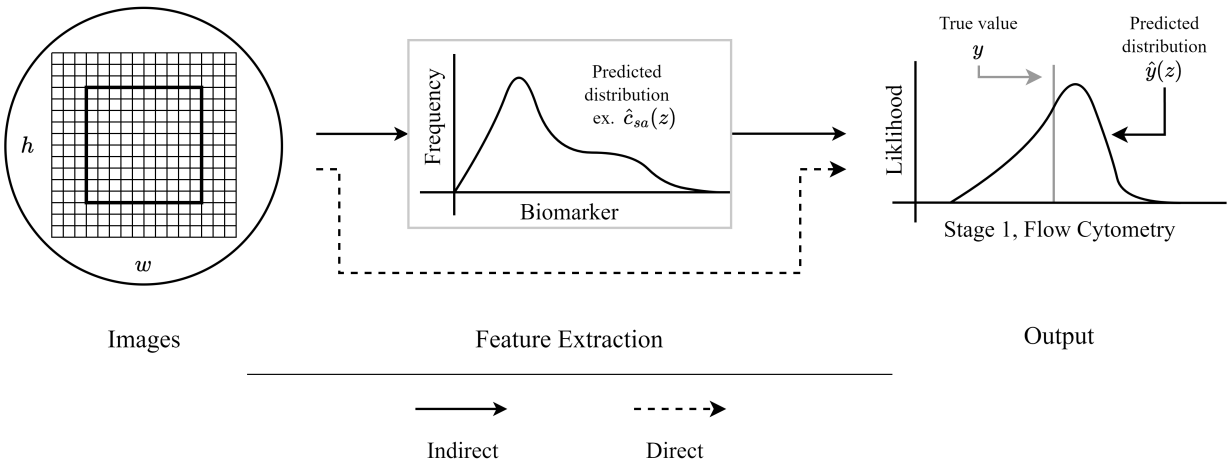



Figure 3.3: Types of considered modelling approaches.

### 3.2.1 Indirect Approaches

For the indirect approach, an estimation of well-empty area distribution is the primary focus of the thesis. Firstly, the creation of summary variables of images, such as well-empty areas using kernel or convolutional filters is performed. Secondly, a comparison of the visual biomarker distributions is performed with the assumption that similar distributions have similar flow cytometry values.

This first approach can be done using non-black-box and black-box approaches. The non-black-box technique uses the idea of image convolutional kernels to provide outputs in the form of 'candidate pixels' which indicate the presence of the phenomena it is intended to capture. An optimizer determines the specifications of these kernels and aims to mimic human-labelled data using a supervised learning formulation. The task is a binary classification task where the outputs from each candidate pixel represent a Boolean value, or true or false to if it is likely present. These Boolean values can be used to create combined or singular metrics to simplify the process for other algorithms. Figure 3.4 shows three of the proposed or implemented techniques.

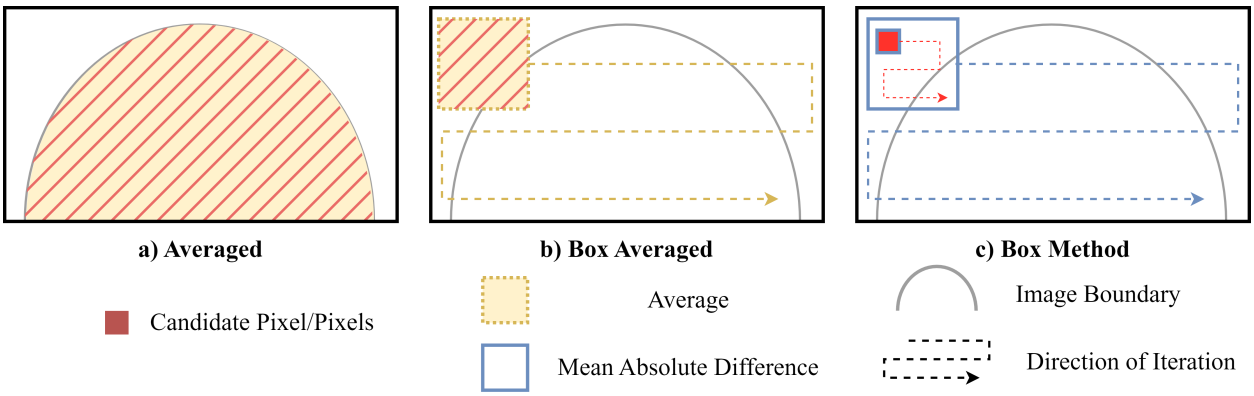



Figure 3.4: Three considered white-box kernels for feature detection.

Averaged outputs a single value across the entire well representing the colour across the entire well. Box-averaged is similar but instead considers subsections of images. For these first two methods, a model is formed between the average pixel brightness, $x$, and the cell surface area with a sigmoid function of the form $f = \frac{a}{1+e^{b(x-c)}}$. These models are trained to have a minimization of the mean squared error (MSE).

The Box method uses the mean absolute difference to find areas where the phenomena are potentially present followed by a smaller box to capture the edges of the area present. These computed values are compared to pre-defined or optimized thresholds with values under the threshold marked as negative and exceeding marked as positive. These kernel parameters can be treated as variables in an optimization problem with the following variables as shown in Table 3.1.

The intuition of all three approaches relies upon empty areas having a darker and smoother texture when compared to covered areas. Methods are ordered in terms of increasing complexity and computation time with each considering smaller subsections of the well. While Averaged considers the overall well colour, Box-averaged considers subsections of the well. Finally, Box has a larger box sized $n_1$ to identify candidate areas of well-empty

Table 3.1: Types of Pre-Defined and Considered White-Box Kernels

| Name | Desc. | Optimization Eqn. |
|---|---|---|
| Averaged | Form a sigmoid regression equation based on average brightness | $\min_{a,b,c}(\hat{f}(a,b,c) - c_{sa})^2$ |
| Box-averaged | Form a sigmoid regression equation for an image split into a square of $n$ side length | $\min_{a,b,c;n}(\hat{f}(a,b,c) - c_{sa})^2$ |
| Box | Count if $\alpha_1$ is met for an area $n_1$ and if $\alpha_2$ is met in an area sized $n_1 \times n_2$ | $\min_{\alpha_1,\alpha_2,n_1,n_2} \lvert\hat{f}(\alpha_1,\alpha_2,n_1,n_2) - c_{sa}\rvert$ |

area with a smaller box $n_2$ used to more accurately capture the borders between the cell surface and the empty area.

These optimization problems have models which are transparent and can therefore be solved using conventional optimization algorithms. Given the discontinuity in the function, a solver such as Nonlinear Optimization by Mesh Adaptive Direct Search (NOMAD) 4 can be used [17]. Using image editing software, masks are drawn for each well which indicates the presence and location of the phenomenon. These masks, which can be split into multiple images to create more data, are used to calculate the percentage of the phenomena in each image as a numerical output. For the first two transparent approaches, a numerical output of the total cell surface area is the output. In the latter, a mask similar to R-CNN is outputted.

Alternatively, a black-box CNN has significantly greater model complexity and capability to approximate complex functions. However, the lack of explainability makes it difficult to identify the true mechanism by which a surrogate model achieves its respective results. The original images are fed through the preprocessing pipeline and used as the input images. This process is shown in Figure 3.5 with a numerical sigmoid output.

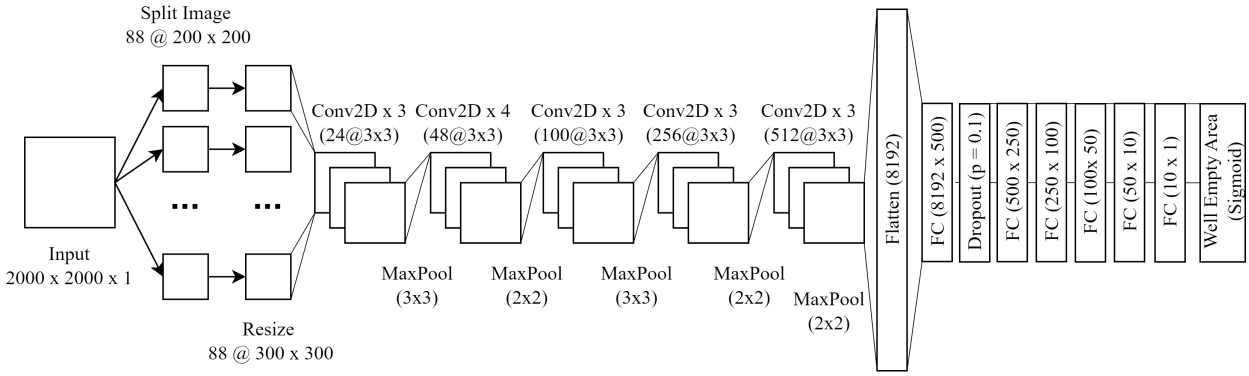



Figure 3.5: Indirect approach with convolutional neural networks

The input images, which are taken from a well as shown in Figure 3.1a are split into a further 88 images. These images are formed from a 10×10 grid with three corner images removed due to the vignetting effect [36]. This split is done to increase the number of unique images available to the deep learning algorithm. Splitting each input image into 88 separate images increases the pool of available data by effectively two orders of magnitude. The design of the CNN is loosely based upon VGG-19 [44] with multiple convolutional layers followed by a rectified linear unit activation (ReLU) and a max-pooling layer. Additionally, this is more representative of the bioprocess as each well, even with the same inputs, is not uniform within each well or between different wells and produces a stochastic output. After the feature extraction of convolutional layers, fully connected layers are connected to a single sigmoidal output to form a regression problem to estimate a well-empty area.

### 3.2.2 Direct Approaches

For the direct approach, a CNN is designed to take in the set of images and estimate the flow cytometry value. This approach requires no additional data labelling beyond the flow cytometry value associated with the image. Each of the images is split into 4 separate pieces to increase the number of samples for training. This split is far less aggressive than the 88 proposed in the indirect approach as it is believed that larger segments of the image are

required to effectively predict flow cytometry values. The structure of the CNN used in this thesis is shown below in Figure 3.6 with the output changed depending on the task.

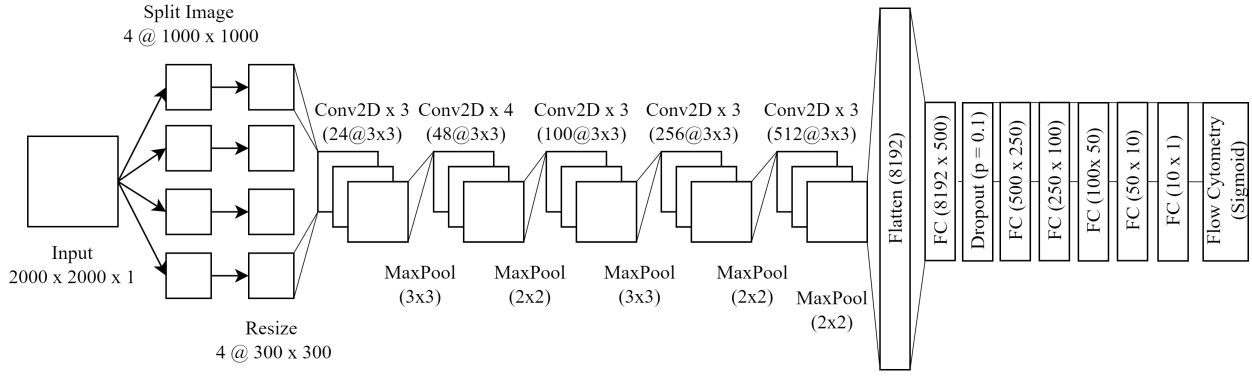



Figure 3.6: Direct approach with a convolutional neural network.

In this formulation, the output layer is a sigmoid layer which forms a regression problem with the output representing the percentage value of the experimental quality. This approach does not require apriori knowledge of the causes of high flow cytometry values which can be visually obtained. However, the black-box nature of deep CNNs means the mechanisms or visual attributes which correlate to high flow cytometry cannot be observed. To ensure the model is not overfitted, a stratified train, validation, and test split are done across wells. Four total wells form the validation and test splits and consist of two wells below and two wells above a 90% stage 1 flow cytometry threshold. This is done to avoid reporting results where well-specific features such as average brightness or blurriness are learned and correlated to flow cytometry values.

Additionally, significant variance is expected in image quality, particularly for low-flow cytometry wells. Low flow cytometry indicates a large proportion of cells have failed to differentiate which, anecdotally, does not occur evenly throughout the well. In contrast, for a high-flow cytometry well, the full well must be consistently high quality as few cells have failed to differentiate. Therefore, it is expected a large amount of images will be misclassified at lower flow cytometries with the distribution of images between low-flow and high-flow

wells likely intersecting. Therefore, while the training is done per quarter well, the prediction of experimental quality must be done with multiple images per well up to a statistically significant sample.

In both formulations, direct and indirect, the CNNs are trained using pytorch [45], a deep-learning framework which allows for the implementation of self-defined architectures. Images are loaded in as a matrix of integers with the numpy [46] and Pillow [47] python libraries. Loss functions are set to MSE to ensure mistakes which are further away are penalized more heavily. Unless otherwise specified, the selected optimizer is Adam [48] with a learning rate $\alpha = 5 \times 10^{-5}$ and default values for the decay $\beta_1 = 0.9, \beta_2 = 0.999$ and epsilon stability $\varepsilon = 10^{-8}$. This is done due to the "robust" performance of optimizers with adaptive learning rates [49] as the focus of the thesis is to identify feasible approaches for stem cell process control. Finally, mini-batches of 1 image are used with weight updates after every 100 images.

# Chapter 4

# Results and Discussion

Images were collected by multiple members of the SCBL. The data which was collected alongside the respective flow cytometry values are summarized in Table 4.1.

Table 4.1: Bioprocess Modelling Collected Datasets

| Batch | Date | Wells | Flow Cytometry | Labeled? |
|---|---|---|---|---|
| 1 | Jun 9th, 2023 | 1 to 3 | - | Y |
| 2 | Jul 24th, 2023 | 4 to 8 | 27.3, 25.2, 50.8, 41.5, 33.1 | Y |
| 3 | Aug 22nd, 2023 | 9 to 10 | 81, 72.7 | N |
| 4 | Sep 12th, 2023 | 11 to 12 | 95.3, 96.3 | N |
| 5 | Jan 3rd, 2024 | 13 to 15 | 97.2, 97.6, 97.2 | N |
| 6 | Jan 14th, 2024 | 16 to 21 | 80.5, 64, 85.5, 85, 72, 71 | N |
| 7 | Jan 18th, 2024 | 22 to 27 | 45, 30, 78, 78, 78, 65, 44 | N |
| 8 | Jan 30th, 2024 | 28 to 30 | 93, 93, 93 | N |
| 9 | Feb 16th, 2024 | 31 to 33 | 98.2, 98.2, 98.2 | N |
| 10 | Feb 17th, 2024 | 34 to 37 | 98, 98, 98 | N |
| 11 | Apr 20th, 2024 | 38 to 43 | 93, 93, 92, 93, 91, 96 | Y |
| *Data Available* | | | | |
| Direct Approach | | | | 12,800 |
| Indirect Approach | | | | 12,144 |

Batch number 1 consists of 144 images each with all other batches consisting of 80 images. Batches of images are generally done on the same plate but on separate wells. Some batches,

notably batch 1, 7, and 11 have variable cell density in each well. In general, the SCBL assumes that wells of identical cell density on the same plate have equal flow cytometry values. This assumption is used in batches 8, 9 and 10 to significantly increase the number of images which we can use for high-flow cytometry wells. In the direct approach, images are split into four with all images consisting of a flow cytometry value used for the classification task for a total of 12,800 images. In the indirect approach, 20 images from batch 1, 59 images from batch 2, and 59 images from batch 10 are labelled. These 138 images can be used as-is for the BBO optimization approach described in Table 3.1. However, the convolutional neural network shown in Figure 3.5 benefits from more separate pieces of training data. Therefore it can be separated into 88 images each for a total of 12,144 images.

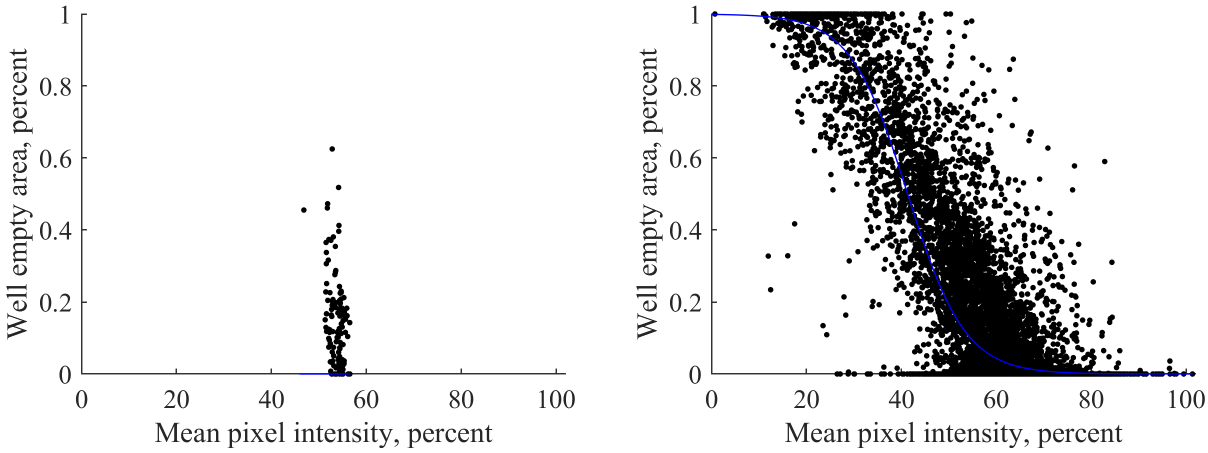
In the results, the capability of each approach to perform feature extraction is explored in Section 4. This is followed by attempts to directly predict flow cytometry values in Section 4.1. In this thesis, a focus is placed on images which are collected approximately 24 hours after initial seeding. All results which are obtained from the wells pertain to data formulated from these images.

## 4.1 Feature Extraction

The results of feature extraction using the transparent approaches described in Table 3.1 and Figure 3.5 are described in the two following subsections.

For the first two models, averaged and Box-averaged, the lack of model complexity and expressiveness results in no train-validation split being performed. An 85% and 15% train-validation split is performed for the Box method due to its greater complexity and is done randomly across all available training images. A sigmoid function in the form $f = \frac{a}{1+e^{b(x-c)}}$

is used to fit the brightness of the image to the cell surface area distribution with the fits shown in Figure 4.1.




(a) Averaged (b) Box-averaged

Figure 4.1: Model of cell surface area extraction with simple models.

In these curves, the average image brightness of an image is typically constant and an ineffective indicator for well-empty area estimation. While line brightness is standardized, this process is imperfect and likely damages this relatively simple approach. This approach is sensitive to the size of the cells which vary the quantity of lines covering the same given area. Smaller cells have greater line density which would increase the average brightness despite covering no additional area. Due to this inconsistency, with a significant number of wells at full coverage, the predictor consistently guesses a value of zero. However, even with this shortcoming in these simple models, a $n = 10$ split for a Box-averaged approach allows for significant predictive power. Therefore, consideration of smaller sections of the images can plausibly increase performance.

The remaining two approaches generate parameters which directly estimate the cell surface area and therefore no curve fit is required. The box approach optimizes the 5 thresh-

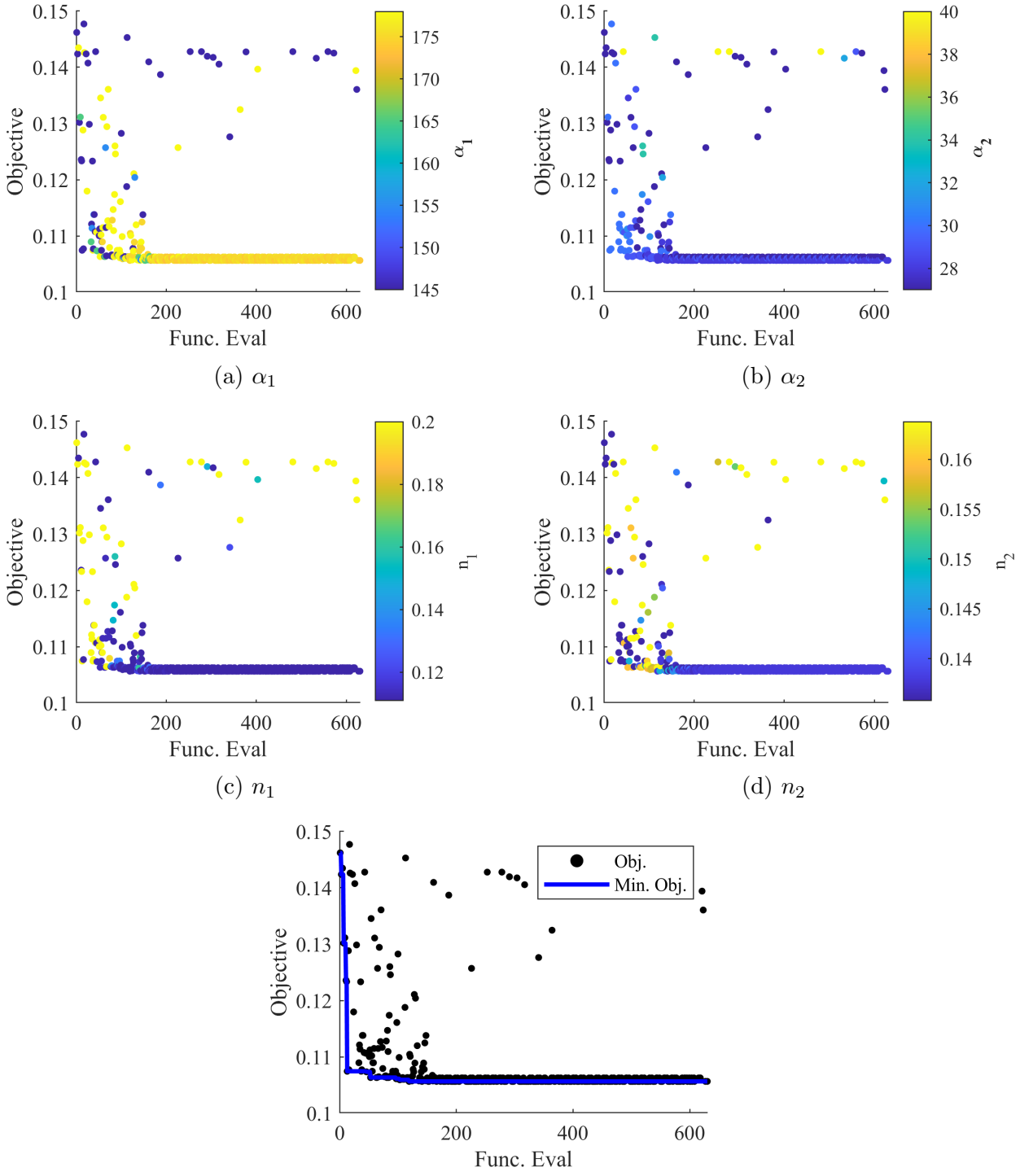
olding variables with NOMAD 4 to find cell surface area. Due to its greater capacity, the selection of the best model is done with a train-validation split of 70% and 30% but results include all images. For the deep CNN, a train, validation, and test split of 90%, 5%, and 5% is used. At each iterate, an overview of the model loss can be plotted and is shown in Figure 4.2 for the Box method and Figure 4.3 for the CNN.

Both curves show a reduction in training loss with increased epochs. In Figure 4.2, the change in variables is demonstrated with the colour of the plots with the position indicating loss. NOMAD 4 quickly converges to a model in under 50 function evaluations with approximately 10% error and finds minimal improvements after this. Variable values fluctuate between the 50 and 200 function evaluations with the optimizer eventually settling upon $\alpha_1 = 174$, $\alpha_2 = 28$, $n_1 = 0.111$, and $n_2 = 0.1378$. Similar observations in the CNN are also observed with few improvements observed after approximately 10,000 minibatches are done.

To compare the overall quality of these approaches, the predicted and actual well-empty areas are plotted for all four models. An optimal model would have all predictions along the diagonal which is shown as a blue line as shown in Figure 4.4. Across all results, the deep CNN outperforms simpler approaches in terms of MSE terms error.

Unsurprisingly, the averaged approach consistently guesses low well-empty area coverages throughout the domain and shows no capability to fit or estimate phenomena. This is primarily due to the average image brightness being too similar for a proper regression. In contrast, the Box-averaged and Box approaches show some capability to fit the model but with significant variances and large amounts of bias for the Box method. In contrast, the variance is greatly reduced with minimal bias for the CNN classifier.

Significant misclassification is still observed with these simpler approaches. While the




(a) $\alpha_1$

(b) $\alpha_2$

(c) $n_1$

(d) $n_2$

(e) Best Objective

Figure 4.2: Box method variables and loss curve of cell surface area model.

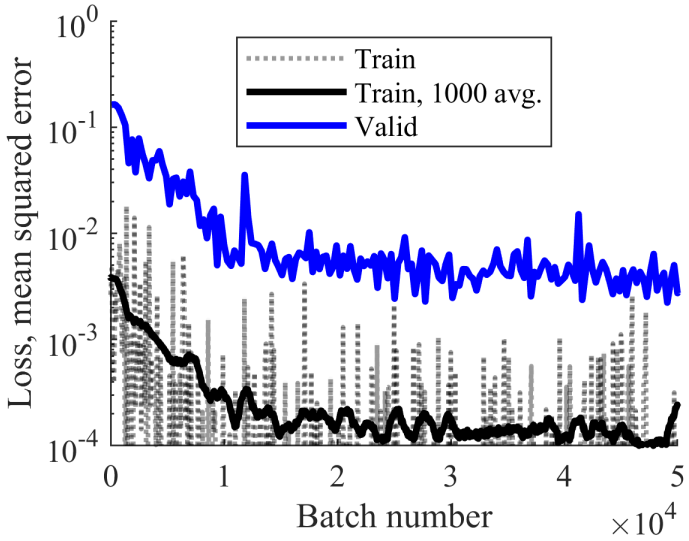
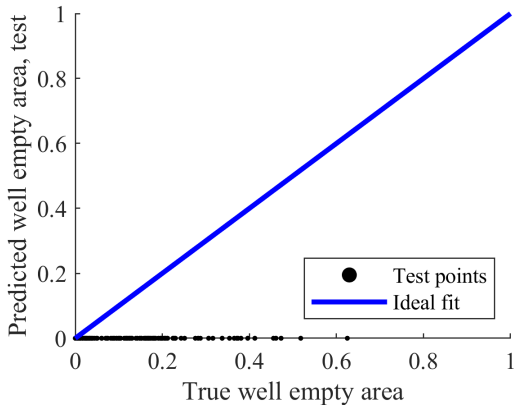


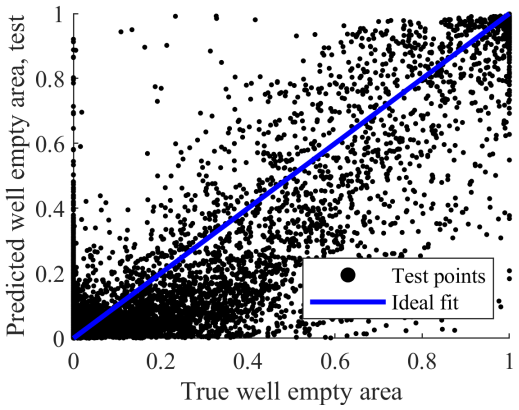

Figure 4.3: CNN loss curve of cell surface area model.

error of the Box method is greater than the averaged approach, the averaged approach has a lower error primarily due to the significantly lower variance of the output label. Many test points with no true well-empty area are predicted to have significant amounts of coverage using the Box-averaged and Box approach. While the error is reduced with the higher capacity approach of the Box-averaged approach, the transparent models have noticeable errors in low well-empty area images. Given the existing literature described in Section 2.1.1 on the importance of cell density and localized aggregation, these types of misclassification are concerning. The only model which possesses sufficient capability to differentiate between high well-emtpy area and low well-empty area is the CNN which is used for the indirect approach in the remainder of the thesis.

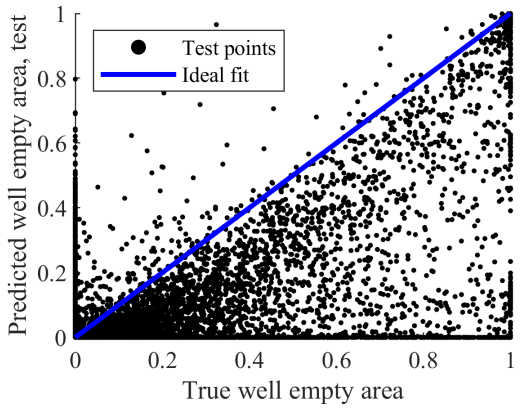
Additionally, the author notes that while on paper, the Box method performs worse than all other methods, it is the only one which predicts the location of target features. While all other approaches have been formulated as regression tasks, the Box method resembles novel ML approaches such as R-CNN through the estimation of a "mask" over the original image as shown in Figure 4.5.



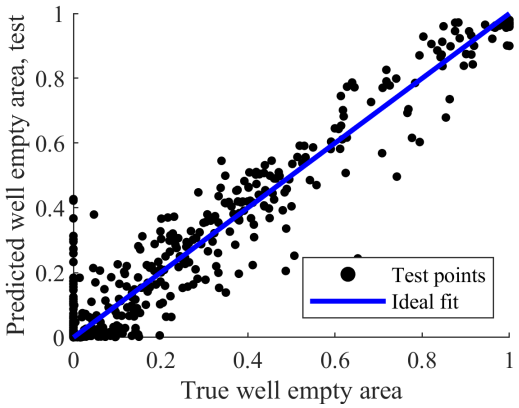

(a) Averaged, MSE: 0.034




(b) Box-averaged, MSE: 0.019



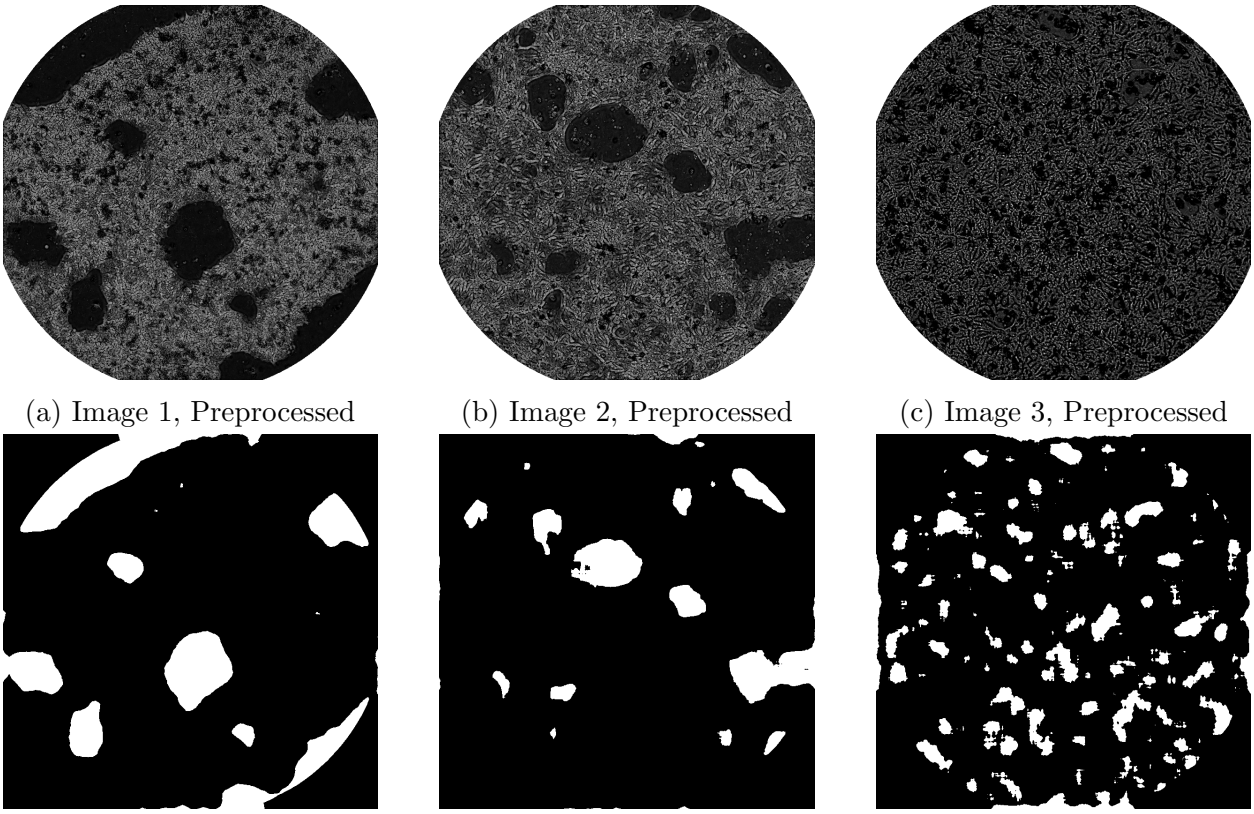

(c) Box method, MSE: 0.042




(d) CNN, MSE: 0.004

Figure 4.4: Fit quality of indirect feature extraction

While consisting of greater error than the Box-averaged approach in Figures 4.4b and 4.4d, the masks still show great capability to identify the location, and approximate size, of well-empty area phenomena. The cause of the increased mean squared error is primarily due to the undersizing of predictions of well-empty area as seen in the top left of Figure 4.5b. These issues can be exacerbated by variances in image brightness or random noise in the image which can be caused by dead cells. In Figure 4.5b, dead cells are notably seen in the larger empty areas close to the center of the well. These dead cells show up as subtle white dots and appear to the simple box classifier to be occupied areas. These errors explain the significant bias of the box classifier with greater model complexity and capacity needed to properly capture the target function.



(a) Image 1, Preprocessed (b) Image 2, Preprocessed (c) Image 3, Preprocessed

(d) Image 1, Predicted (e) Image 2, Predicted (f) Image 3, Predicted

Figure 4.5: Box method prediction of well-empty area phenomenon.

These changes between the images demonstrate a rationale for ML approaches. Even with the brightness standardization applied in the preprocessing, the brightness of the three input images is still noticeable. While the preprocessing pipeline could be further refined to improve the brightness issue or additional capacity could be added to remove these spots before using the Box method, fixing all issues through manual intervention is impractical. For this reason, black-box CNN models, which have significantly lower MSE, allow for the potential mitigation of these future issues without specific design intent. This is achieved through the varied feature extraction it performs and ensuring the model does not excessively overfit through the loss curve in Figure 4.3.

## 4.2 Flow Cytometry Prediction

Flow cytometry prediction accomplished through the indirect approach necessitates similarity in well-empty area distributions or a direct model with low enough loss or error. Using the deep CNNs the well-empty area distributions are compared in Section 4.2 and the capability of the direct approach is discussed in Section 4.2.1.

### 4.2.1 Comparison of Well-Empty area Distributions

All well-empty area distributions are shown in Figure 4.6 as probability density functions. Each distribution is made up of the 7040 images which make up each well and was inferred by the CNN model. It is noted that while some of these images were part of the train set, the comparison of well-empty area distributions is not related to the issue of overfitting. The x-axis of each distribution represents the well-empty area with 1 indicating a fully empty square with no cell coverage. The opposite, a 0, indicates full cell coverage with no empty area. Additionally, the y-axis is logarithmic across four orders of magnitude and represents the relative normalized frequency. Due to the wide range of magnitudes it covers, the author notes that small differences in the height of the discretized bars can indicate significant differences in frequency.

In these distributions, some trends are observed which suggest the similarity of certain wells. For low-flow cytometry wells, the distribution generally sees a greater frequency of fully empty squares, or the well-empty area is equal to one. This leads to a characteristic "bowl" shape where the majority of the wells are either fully covered or fully empty of cells. In contrast, many high-flow wells tend to have high cell coverage with minimal well-empty areas. This leads to a distribution with an exponentially decaying shape.

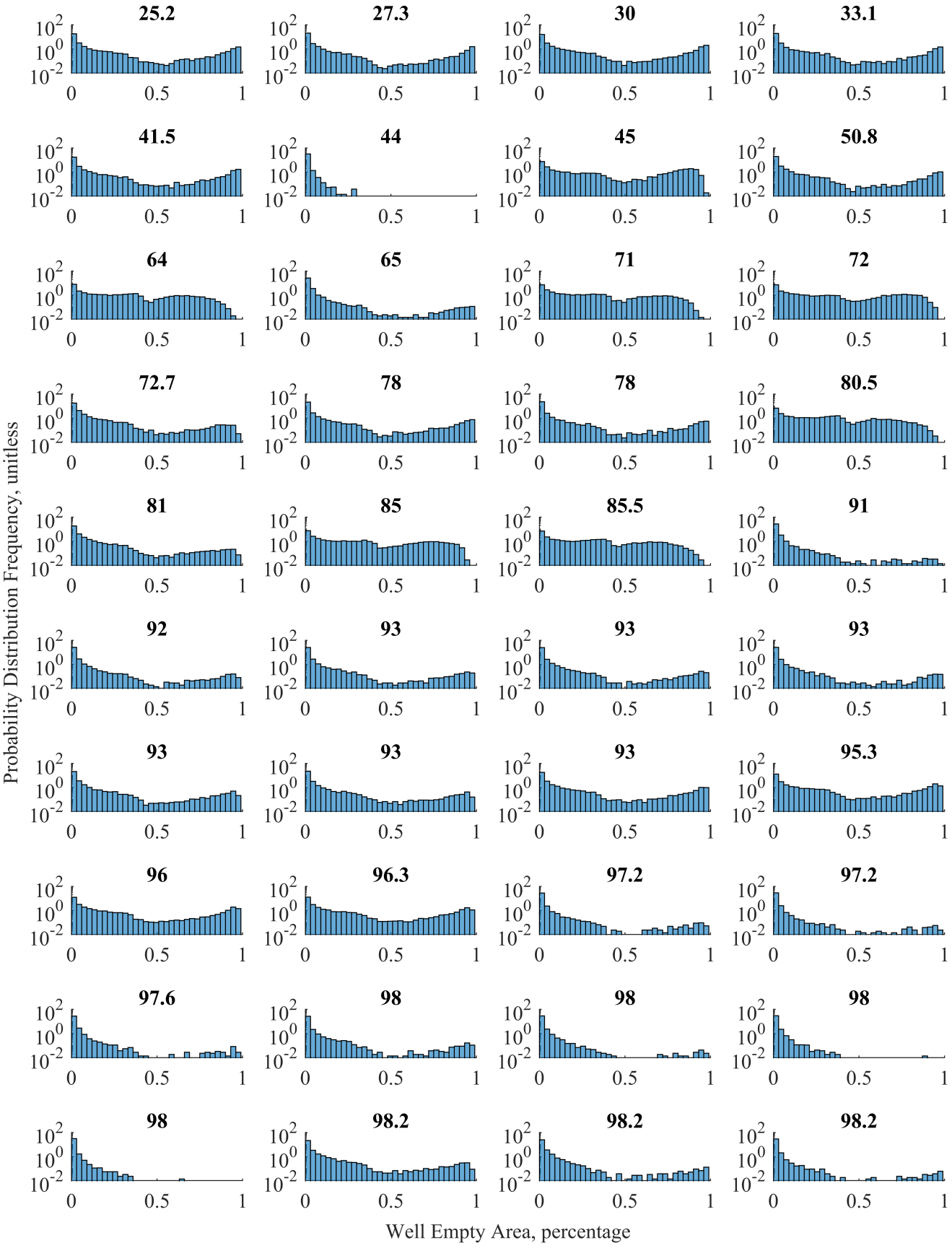



Figure 4.6: Well-empty area distributions for all flow cytometry.

While no obvious markers to determine flow cytometry values are visible, the author notes that distributions are very similar for plates which were seeded with the same cell density and same flow cytometry values. All wells from Batch 2, namely 25.2%, 27.3%, 33.1% 41.5% and 50.8%, share the characteristic low flow cytometry "bowl" shape with a dip in the distribution around the 0.5 mark. Batch 10, made up of 3 wells at 98% flow cytometry shows very similar exponential decay distributions with a very low frequency of fully occupied cells. Even in cases where the distribution does not match the expected result, this similarity is present. For example, Batch 4, is made up of two wells at 95.3% and 96.3%, while appearing closer to a bowl than the characteristic exponential decay, both have a downward dip in their distributions at about the 0.95 mark. The author is unable to propose a biological explanation for this phenomenon but notes the observation.

Despite these broad generalizations, significant overlap and noise make classification by inspection very difficult. The wells which have flow cytometry values of 95.3%, 96%, and 96.3% have "well" shapes and would be difficult to separate from low-flow wells. Similarly, the wells with flow cytometry values of 44% and 65% appear similar to high-flow wells. The author proposes these well-empty area distributions are correlated to flow cytometry but insufficient on their own to form a sufficient surrogate model.

To quantify the similarity, or lack thereof, between distributions and an earth-mover metric, Wasserstein distance is used from the Python library scipy [50]. This function, wasserstein_distance_nd, shows the relative "edit" distance between two distributions. Figure 4.7 shows the relative "distance" in a colour bar for all wells ordered by flow cytometry. The diagonal from the lower-left corner to the upper-right corner has zero difference as the distribution it is compared to is itself.

The Wasserstein metric shows minimal capability to define a flow cytometry function as

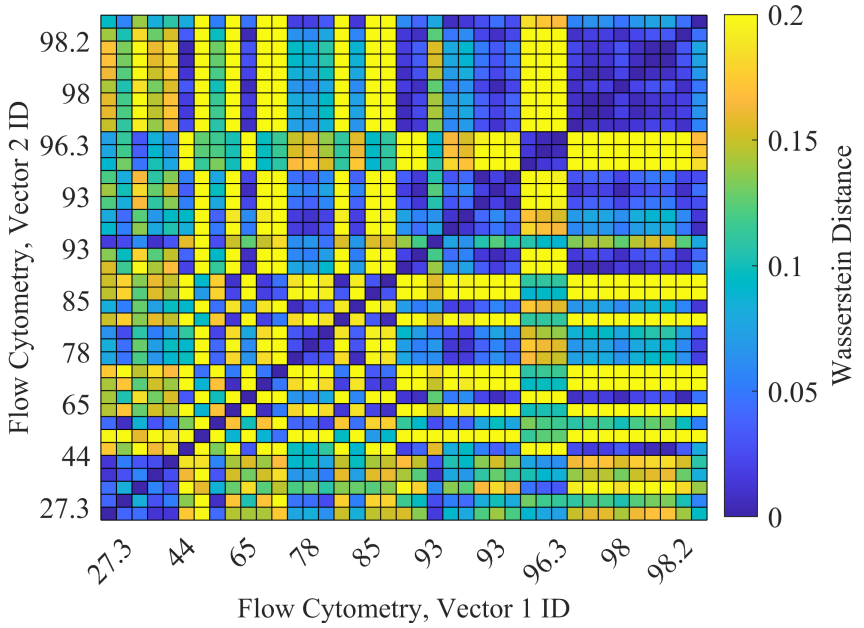


Figure 4.7: Wasserstein distance of well-empty area distributions.

the similarity between wells is inconsistent. Some patches along the diagonal are present, notably below 44% flow cytometry, between 78-81%, and above 96% are seen. However, even with these patches, along their respective rows and columns, the colouring does not uniformly shift to a greater Wasserstein distance with changes in flow cytometry as would be required for proper classification. Of particular note is the comparison of very high flow cytometry wells which are above $\geq$98% to the wells around the 96.3% mark which show great dissimilarity and greater similarity to the low flow cytometry wells below 44%.

While it could be argued that there is a trend of the Wasserstein metric to increase "on average" with changes in flow cytometry, this only suggests the phenomenon of well-empty area distribution is correlated to flow cytometry. Therefore, it is insufficient as a predictor for flow cytometry as the variance is too significant and the well-empty area distributions do not describe enough about the system to properly model the system. The aforementioned phenomena of experiments from the same batch with similar flow cytometry values having

similar distributions are reinforced in the Wasserstein metric. For example, batch 2 is defined by the rows and column numbers 1, 2, 4, 5 and 8 which show significant similarity. This is indicated by the bluer colour that these rows and columns share. Notably, row and column numbers 6 and 7 of batch 7 have similar flow cytometry values but a significantly different well-empty area distribution.

Additionally, batch 11 in Table 4.1 is noted to be one of two batches with variable cell density albeit all flow cytometry values are similar. Therefore, this suggests the argument made by Gage et al. [18] and Toyoda et al. [20] of higher cell density leading to higher flow cytometries may not apply beyond the utilized cell densities. Therefore, the considered cell density values of the existing literature are likely lower than those of this thesis and therefore the correlations do not hold. This suggests the factors which affect differentiation above a certain cell density are still affected by an unspecified and as-of-yet undetermined phenomenon.

### 4.2.2 Direct Approach

Given the lack of clarity in the effects and phenomena which drive low and high flow cytometry at sufficiently high cell densities, a black box and opaque model is warranted. This is achieved through the aforementioned training of a CNN which directly predicts flow cytometry. Training is done with the initial settings described in Section 3.2 which generates the train/validation loss curve as shown in Figure 4.8

A gradual reduction in train and validation loss indicates a properly fit model. Although the validation loss is significantly higher than the train loss, no obvious signs of overfitting are present with the validation loss not increasing over time. For a mean squared error of 0.02, this suggests an average error of 14% for each image which may be a feasible model

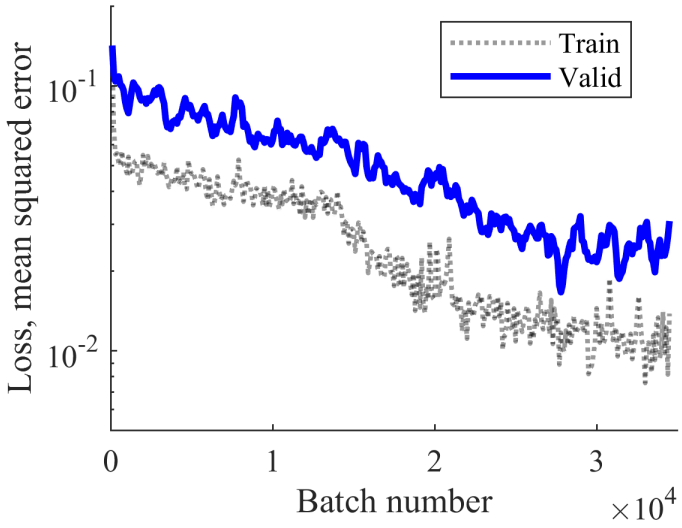


Figure 4.8: CNN loss curve of direct model.

if the errors on each set of the wells are somewhat normally distributed. This plausibility suggests the feasibility of the approach and therefore a hyperparameter sweep is warranted. A random sampling of the learning rate $\alpha$ and dropout $p$ parameters is performed to find regions in the parameters where loss is minimized as shown in Figure 4.9.

Relatively low learning rates under $10^{-4}$ and dropout values below 0.15 show the lowest validation losses in the validation test split. The stratified k-fold (k=5) is set up according to Table 4.2 with a learning rate of $\alpha = 5 \times 10^{-5}$ and dropout $p = 0.125$. While a hyperparameter sweep along each fold could further improve accuracy, this was not done due to the excessive computational cost this would incur. Finally, the k-fold split is stratified along for those under and above 90% flow cytometry. The distribution of the image predictions from the test set from each of the five folds is shown in Figure 4.10.

A further analysis of these results is done by separating the labels into two classes. Good flow cytometry wells are defined as those with over a 75% stage 1 flow cytometry value with all other wells defined as poor. This differs from the stratified split as shown in Table 4.2. While the flow cytometry percentage for a stratified split was selected for data considera-

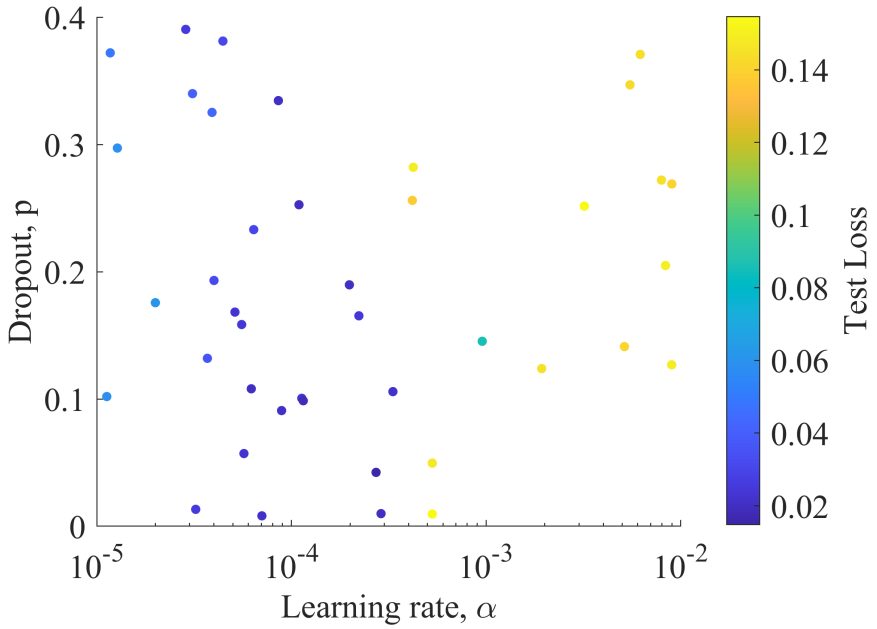



Figure 4.9: Hyperparameter sweep of direct CNN.

tions and to ensure even classes, the analysis is discussed from the perspective of its potential implications in a quality control process. In this context, wells close to but below a 90% can arguably be considered acceptable and good.

The errors of the model are distributed unevenly in the domain. Poor flow cytometry images have an average error of 12.9% with good flow cytometry images having a significantly lower 5.5%. These errors result in an average prediction error of 10.2% for poor

Table 4.2: K-Folds Splits for Direct Approach

| Fold | ≥90% | ≤90% | Flow Values | Val. Loss |
|---|---|---|---|---|
| 1 | 13, 40 | 5, 18 | 97.2, 92, 25.2, 85.5 | 0.0209 |
| 2 | 38, 41 | 6, 11 | 93, 93, 50.8, 41.5 | 0.0189 |
| 3 | 17, 37 | 4, 10 | 64, 98, 27.3, 72.7 | 0.0213 |
| 4 | 42, 43 | 8, 9 | 91, 96, 33.1, 81 | 0.0088 |
| 5 | 11, 33 | 16, 20 | 95.3, 98.2, 80.5, 72 | 0.0132 |
| Average | | | | 0.0166 |

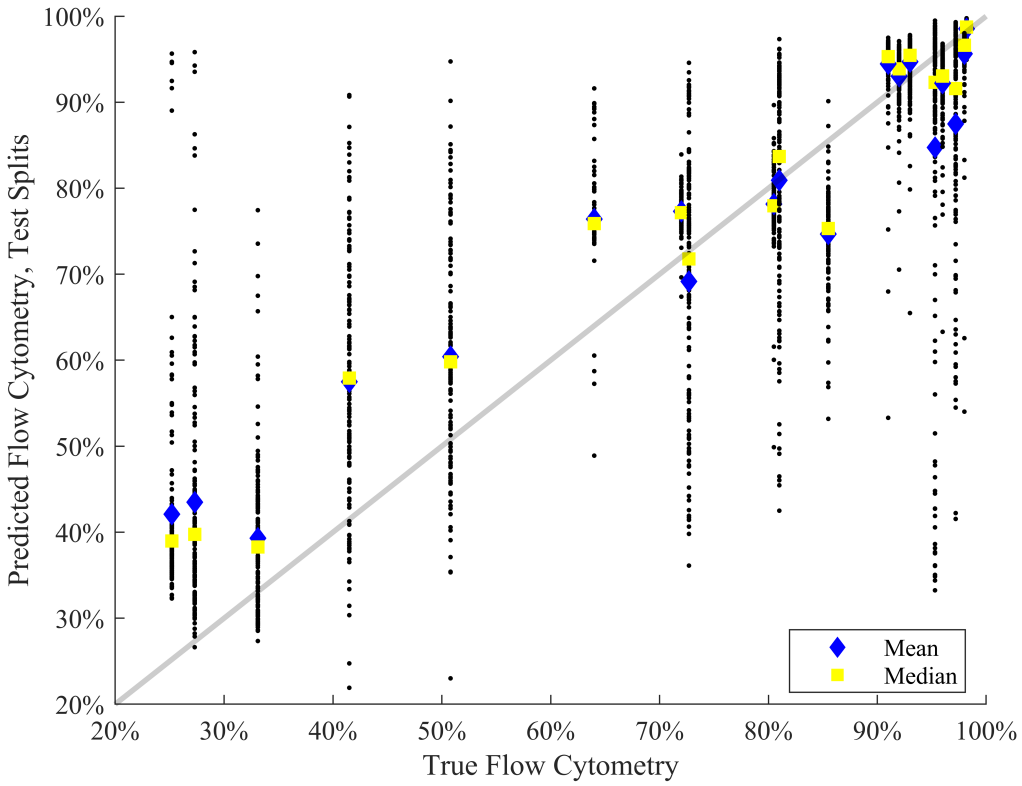



Figure 4.10: Predictions of flow cytometry with direct approach.

flow cytometry wells and -2.6% for high-flow cytometry wells, indicating bias for poor flow cytometry images and variance in good flow cytometry variance. This error distribution is directly observed as a result of the greater variance observed in poor flow cytometry wells as suggested in Section 3.2.1. Images from poor flow cytometries have greater variances in the predicted value from the CNN surrogate. Wells under 75% flow cytometry have a mean standard deviation of 10.4% compared to 6.7% for wells above 75% flow cytometry.

Broadly speaking, the mean and median of the distributions correlate strongly towards the ideal grey line. The average of the mean distribution error is 7.0% across all wells. The poor flow and good flow cytometry distributional errors are 10.8% and 4.2% showing the errors of the poor flow cytometry wells are not overcome with greater sampling of the well.

The error of the median prediction shows slightly improved results with average errors of 9.4% and 3.4% respectively. This suggests an average image of the well, which could be considered "representative" could have a significant correlation to the flow cytometry value. However, defining the "median" image may prove difficult in practice without first obtaining the full distribution.

### 4.2.3 Analysis of Errors

While significantly improved results are seen in the direct approach when compared to the indirect approach, few explanations for the predictions are available due to the opaque nature of the model. For this reason, an analysis of errors is performed to identify the situations in which misclassification occurs and to highlight possible weaknesses of the classifier. Images and predictions are classified into three categories: 1) high or above 90%, 2) medium or between 50% and 90%, and 3) low or below 50%. These three categories are selected as differentiations above 90% represent a sufficient quality differentiation, between 50% to 90% representing the typical performance of the SCBL, and below 50% representing a significant defect was present and likely observable. The numbering of each group is shown in 9 groups in Table 4.3.

The higher number of categories in Section 4.2.2 is intended to draw emphasis upon errors where significant misclassification occurs as opposed to a "typical" or "average" error as described in the prior section.

This table resembles a confusion matrix and is used to identify the types of misclassification in frequency and highlight situations in which they occur. For a perfect classifier, Classes 1, 5, and 9 would have non-zero values with all remaining classes being zero. Of

Table 4.3: Group Numbering of Classification Types

| | | $y$ | | |
|---|---|---|---|---|
| | | Low | Med | High |
| $\hat{y}$ | Low | 1 | 2 | 3 |
| | Med | 4 | 5 | 6 |
| | High | 7 | 8 | 9 |

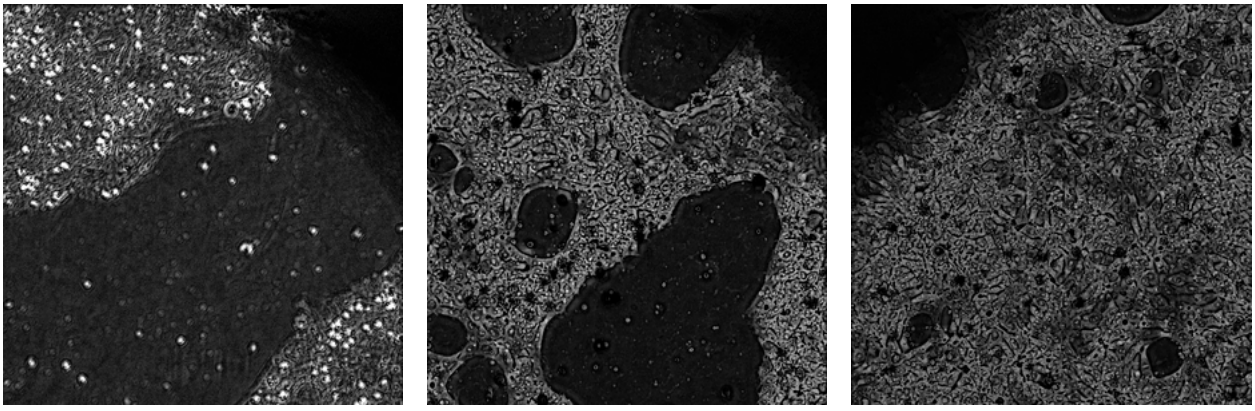
primary concern to the author is errors where significant misclassification occurs, namely Classes 3 and 7. The "confusion matrix" results from the first k-fold are presented in Table 4.4.

Table 4.4: Confusion Matrix of Direct Method

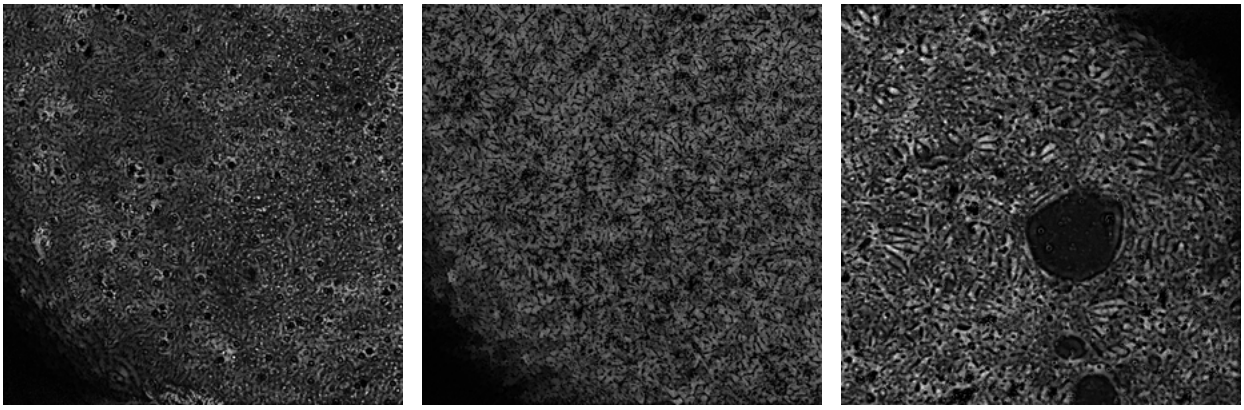
| | | $y$ | | |
|---|---|---|---|---|
| | | Low | Med | High |
| $\hat{y}$ | Low | 149 | 14 | 3 |
| | Med | 1 | 153 | 0 |
| | High | 1 | 109 | 211 |

Using this matrix, out of all 641 images in the first test k-fold, 513 images are correctly "classified" with 128 incorrectly classified. The significant majority of errors involve misclassifying medium flow cytometry images as high flow images. Errors of most particular concern, Classes 3 and 7 are at a respective 4 images total. A visual analysis of these errors can be performed by comparing images from Classes 3 and 7 to Classes 1 and 9 where the classifier and base values agree. Figure 4.11 shows all images from Classes 3 and 7 with two low-flow and 1 high-flow images.

Anecdotally, and through observation in Figure 4.11, low flow cytometry images such as in Figures 4.11a and 4.11b have irregularly shaped and large gaps in contrast to the regular uniformly distributed nature of high flow image in Figure 4.11c. In the Class 3 misclassifications shown in Figures 4.11d and 4.11e, the well has relatively high cell coverage, uniform-sized cells, and few gaps are noticeable and resembles the high-flow image of Figure



(a) Class 1, $\hat{f}$ : 38.6%, $f$ : 25.2% (b) Class 1, $\hat{f}$ : 35.8%, $f$ : 25.2% (c) Class 9, $\hat{f}$ : 97.2%, $f$ : 97.1%



(d) Class 3, $\hat{f}$ : 91.4%, $f$ : 25.2% (e) Class 3, $\hat{f}$ : 96.1%, $f$ : 25.2% (f) Class 7, $\hat{f}$ : 45.1%, $f$ : 97.2%

Figure 4.11: Comparison of preprocessed images and error type.

4.11c closer than that of its low-flow cytometry counterparts. In contrast, the high flow cytometry image of 4.11f, consists of two larger gaps suggesting to a human observer it could be from a low flow cytometry image.

Given the inconsistencies with the well-empty area discussed in Section 4.2 and the difficulty in correctly classifying these images, the author suggests the problem has a high Bayes error. The overlap between image distributions, while hard to quantify from just designing a CNN, is likely significant. As noted in the motivation, or Section 1.1.3, the significant variance in experimental pancreatic stem cell differentiation means the problem is highly stochastic, or random. Therefore, it is unrealistic and would be suspicious for a

classifier to propose significant capability to define flow cytometry values from small image-based samples of the well.

## 4.3 Future Work

While deep CNNs show promise in modelling the opaque mechanisms which lead to high pancreatic flow cytometry values, future work remains to establish it as an effective methodology. Model improvements could be explored through new ML formulations and increasing methodology maturity through wider implementation. Additionally, the application of the model includes usage in an optimization context and attempting the extension of similar models to other organoids.

### 4.3.1 Hybrid Approaches

One form of model that was not tested in the thesis is referred to as a "hybrid approach" which combines the direct and indirect approaches. This combination utilizes the feature extraction either as an input or an output to a direct method to increase model performance. Two hybrid approaches are proposed; either one where the features are used as an input to assist prediction or where the CNN learns additional features to improve generalization as shown in Figures 4.12 and 4.13 respectively.

In Figure 4.12, the increased feature set is intended to increase the number of inputs to the predictive algorithm. This resembles the stacked neural network (SNN) formulation as proposed by Mohammadi and Das [51]. While CNNs identify higher-level features [41] and feature engineering is not common with deep learning techniques [52]; the cell surface area and density have previously shown potential correlation to flow cytometry in Section 2.1.1. Therefore, this stacked formulation is proposed to provide a more explicit set of features for the surrogate model.

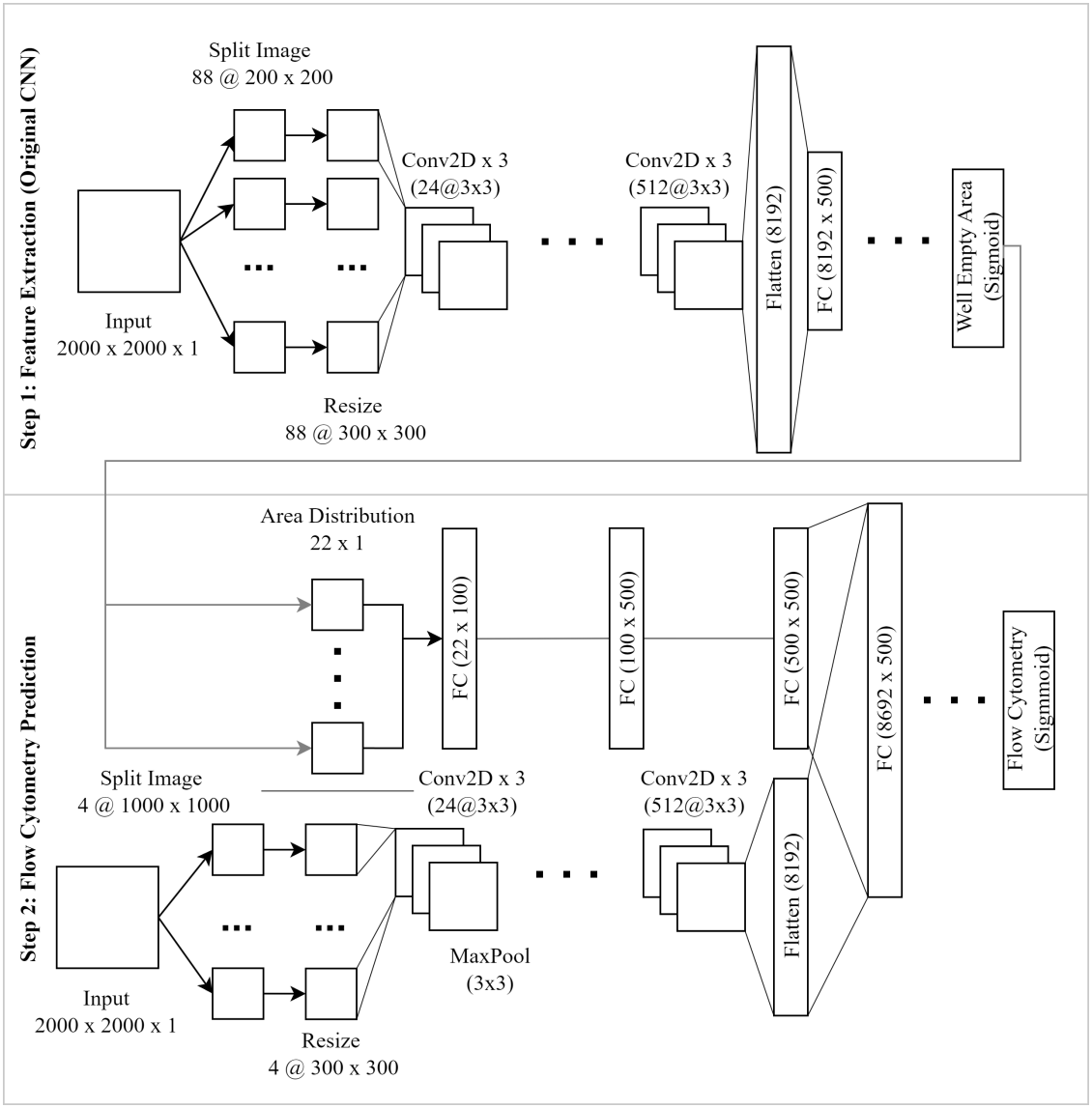



Figure 4.12: Hybrid approach using stacked neural networks.

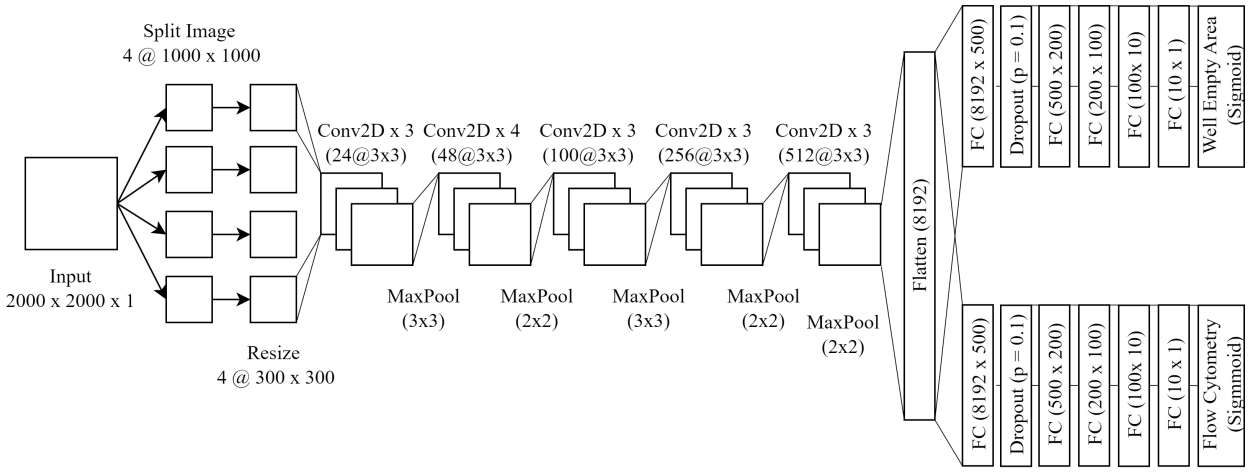



Figure 4.13: Hybrid approach using multi-task formulation.

In contrast, Figure 4.13 involves separate predictions for the biomarker with a multi-task formulation to improve feature generalization [53]. As the convolutional layers in a CNN act as feature extractors, this formulation posits designing the feature set to predict multiple labelled values improves overall regression and classification quality.

### 4.3.2 Experimental Aid and Timing

By using the existing model or an improved variant of said model to estimate the flow cytometry at each step, various assistive approaches to improve differentiation protocols could be achieved.

The modelling of images to differentiation quality can be extended to define the prescriptiveness [54], or the information gained at each step, to quantify the improvement of the differentiation prediction at each stage. Earlier predictions of unsuccessful differentiations can minimize unnecessary expenditures of growth factors or human factors. Each set of images taken at one of these fixed time frames can be formed as an independent problem.

When all input variables are kept constant, it can be reformulated as a stochastic (or uncertain) contextual optimization problem with a cost function $C$. The cost function could be defined as a ratio between the financial or temporal costs relative to successful differentiation percentage minimized given an action z (where 0 and 1 represent stopping and continuing the differentiation at the current stage with its information) shown in Equations (4.1a), (4.1b), and (4.1c)

$$z_{\text{seed}_{24}} \in \min_{z\in[0,1]} \mathbb{E}_{\Theta}\left[c_{\Theta}\left(z; I_{\text{seed}_{24}}\right)\right] \tag{4.1a}$$

$$z_{\text{seed}_{48}} \in \min_{z\in[0,1]} \mathbb{E}_{\Theta}\left[c_{\Theta}\left(z; I_{\text{seed}_{48}}, I_{\text{seed}_{24}}\right)\right] \tag{4.1b}$$

$$\dots$$

$$z_{\text{s}_3} \in \min_{z\in[0,1]} \mathbb{E}_{\Theta}\left[c_{\Theta}\left(z; I_{\text{s}_3}, I_{\text{s}_2}, I_{\text{s}_1}, I_{\text{seed}_{48}}, I_{\text{seed}_{24}}\right)\right]. \tag{4.1c}$$

### 4.3.3 Growth Factor Optimization

Optimization of the growth factors can occur by comparing the value of the new growth factors to the expectation after the initial seeding stage using the initial bioprocess. This expectation can be formed by a model of the original process which would predict the flow cytometry value using the existing methodology. Finally, the actual measured flow cytometry value defines the relative improvement or harm as a differential, $\delta$. Equation (4.2) shows this hypothetical for a model which predicts the Stage 4 flow cytometry given the 24-hour seeding images.

$$\delta = \bar{f}_{\text{s}_4} - \hat{f}_{\text{s}_4}(I_{\text{seed}_{24}}). \tag{4.2}$$

By comparing the experimentally obtained flow cytometry values to the expected values,

the impact of growth factors is theoretically isolated from the well-empty area and other impacts captured in the images. Using the methodology in this way makes two assumptions. Firstly, it assumes that the cells which fail to differentiate at any stage are unable to further differentiate or succeed in later stages; although in certain conditions, multiple differentiation pathways have been recently observed for cell differentiation [7, 27].

Secondly, for modelling, stages of differentiation are assumed to be sequential but independent. These two assumptions allow for the objective of the optimization to be a maximization of the differentiation rate at each stage. With sufficient experimentation, this methodology could quantify the mean difference and variance from the trained surrogate that defines a baseline differentiation protocol. However, the number of experiments required to reach a statistically significant improvement in cell differentiation while varying growth factor input variables is unknown and has not been estimated.

This relationship can be used to define the expectation $E$ of the differentiation with a constant set of growth factors at each stage. Impacts due to the changing of input variables can be compared to this baseline to quantify improvement. This can be done by maximizing the differential, $\delta$, as summarized in Equation (4.3)

$$\delta\left(y\right) = \bar{f}_{\mathrm{s}_4} - \hat{f}_{\mathrm{s}_4}(I_{\mathrm{seed}_{48}}). \tag{4.3}$$

The selection of growth factors, an optimization problem, will require biological intuition and take advantage of prior knowledge with the limited number of experiments. However, with the limited number of variables at each stage, a black box optimization problem could be implemented at each stage to find growth factor combinations which surpass the expectation or baseline experiment.

### 4.3.4 Other Bioprocesses

The proposed methodology could be applied to other organoid bioprocesses if they were able to follow the image-based data collection process described in this thesis. This could include cerebral (brain) and hepatic (liver) organoids which are currently under investigation by the SCBL. Direct deep CNN models can be designed and used to predict flow cytometry values with a similar quantity of data if specific biomarkers are identifiable. However, the author acknowledges differences in timings and growth factors could lead to variances in both the quality and effectiveness of similar black box models for different bioprocesses.

# Chapter 5

# Conclusion

In this thesis, various modelling techniques ranging from simple regression, transparent kernels, and deep convolutional networks are explored to estimate flow cytometry. The successful modelling of flow cytometry would allow for proper quality control in pancreatic stem cell differentiations which at present have high variability.

A CNN which directly models the flow cytometry value from the input images is shown to be the only approach which can approximate the flow cytometry values. While still possessing significant bias in modelling low-flow cytometry values, the average prediction shows the capability to separate these wells from high-flow cytometry values and provide a proper numerical estimate of high-flow cytometry wells. This suggests further research should be performed both in improving opaque models to these stochastic bioprocesses and applying them to optimization loops to further improve its technology readiness.

The difficulties encountered in this thesis are not novel issues and have been thoroughly described by the existing literature. These are primarily a result of the nature of the bioprocess which is high cost, has low data quantities, and high variability. This limits the

approaches both to model and optimize the pancreatic stem cell bioprocess. Additionally, the lack of explainability in the mechanisms of pancreatic differentiation necessitates the usage of black box opaque models to ensure a functional model. While the high cost and low data quantities are partially tackled in the proposed methodology through the splitting of images, this is an imperfect solution and further analysis should be performed to determine its true performance and improvement.

The high variability has not been fully addressed in the thesis and future studies should focus upon these issues. Image quality is highly variable and the effect of brightness and contrast on model performance has not been explored. While these changes in image quality would be likely severely detrimental to simple and transparent models, black box models may show more resistance through proper training and image preprocessing. Finally, better modelling techniques to acknowledge the distribution of flow cytometries as opposed to treating it as separate individual estimation tasks as was shown in this thesis.

Additionally, the generalizability of the model has not been confirmed due to the application of this methodology having been limited to one bioprocess, for one organoid, on one cell line, and only for one step of the bioprocess between the two earliest stages. Deep neural networks are universal function approximators which have a strong tendency to overfit and even with the robust train/validation/test split performed with a k-folds split, the potential limitation of this methodology to this bioprocess cannot be ruled out. Therefore, building a generalizable model with multiple bioprocesses or multiple steps, while computationally expensive, may allow for the finding of more generalizable features which would improve overall model quality and generality.

Despite these shortcomings, the demonstrated regressive capability for predicting flow

cytometry of pancreatic organoids indicates further research should be performed. The thesis proposes variances to both the model and experimental work to improve upon future models with the hope of improving future quality control. While bioprocesses are not directly observable, unlike mechanical engineering manufacturing processes, the modelling of these approaches is the first step to allow for true automated control. Without a proper estimation of the present feasibility of the component, stem cell or not, no informed decisions can be made.

# Bibliography

[1] C. Ramond, B. S. Beydag-Tasoz, A. Azad, M. van de Bunt, M. B. K. Petersen, N. L. Beer, N. Glaser, C. Berthault, A. L. Gloyn, M. Hansson, M. I. McCarthy, C. Honore, A. Grapin-Botton, and R. Scharfmann, "Understanding human fetal pancreas development using subpopulation sorting, rna sequencing and single-cell profiling," vol. 145, no. 16. doi: 10.1242/dev.165480. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30042179

[2] J. Casamitjana, E. Espinet, and M. Rovira, "Pancreatic organoids for regenerative medicine and cancer research," vol. 10, p. 886153. doi: 10.3389/fcell.2022.886153. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/35592251

[3] J. Vives and L. Batlle-Morera, "The challenge of developing human 3d organoids into medicines," vol. 11, no. 1, p. 72. doi: 10.1186/s13287-020-1586-1. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/32127036

[4] P. Gupta, P. A. Perez-Mancera, H. Kocher, A. Nisbet, G. Schettino, and E. G. Velliou, "A novel scaffold-based hybrid multicellular model for pancreatic ductal adenocarcinoma-toward a better mimicry of the in vivo tumor microenvironment," vol. 8, p. 290. doi: 10.3389/fbioe.2020.00290. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/32391339

[5] P. Augsornworawat, K. G. Maxwell, L. Velazco-Cruz, and J. R. Millman, "Single-cell

transcriptome profiling reveals beta cell maturation in stem cell-derived islets after transplantation," vol. 32, no. 8, p. 108067. doi: 10.1016/j.celrep.2020.108067. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/32846125

[6] N. Sharon, J. Vanderhooft, J. Straubhaar, J. Mueller, R. Chawla, Q. Zhou, E. N. Engquist, C. Trapnell, D. K. Gifford, and D. A. Melton, "Wnt signaling separates the progenitor and endocrine compartments during pancreas development," vol. 27, no. 8, pp. 2281–2291 e5. doi: 10.1016/j.celrep.2019.04.083. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31116975

[7] M. B. K. Petersen, A. Azad, C. Ingvorsen, K. Hess, M. Hansson, A. Grapin-Botton, and C. Honore, "Single-cell gene expression analysis of a human esc model of pancreatic endocrine development reveals different paths to beta-cell differentiation," vol. 9, no. 4, pp. 1246–1261. doi: 10.1016/j.stemcr.2017.08.009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28919263

[8] H. Wang, P. Maechler, B. Ritz-Laser, K. A. Hagenfeldt, H. Ishihara, J. Philippe, and C. B. Wollheim, "Pdx1 level defines pancreatic gene expression pattern and cell lineage differentiation," vol. 276, no. 27, pp. 25 279–86. doi: 10.1074/jbc.M101233200. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/11309388

[9] T. Gao, B. McKenna, C. Li, M. Reichert, J. Nguyen, T. Singh, C. Yang, A. Pannikar, N. Doliba, T. Zhang, D. A. Stoffers, H. Edlund, F. Matschinsky, R. Stein, and B. Z. Stanger, "Pdx1 maintains beta cell identity and function by repressing an alpha cell program," vol. 19, no. 2, pp. 259–71. doi: 10.1016/j.cmet.2013.12.002. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24506867

[10] A. Veres, A. L. Faust, H. L. Bushnell, E. N. Engquist, J. H. Kenty, G. Harb, Y. C. Poh, E. Sintov, M. Gurtler, F. W. Pagliuca, Q. P. Peterson, and D. A.

Melton, "Charting cellular identity during human in vitro beta-cell differentiation," vol. 569, no. 7756, pp. 368–373. doi: 10.1038/s41586-019-1168-5. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31068696

[11] R. Tran, C. Moraes, and C. A. Hoesli, "Controlled clustering enhances pdx1 and nkx6.1 expression in pancreatic endoderm cells derived from pluripotent stem cells," vol. 10, no. 1, p. 1190. doi: 10.1038/s41598-020-57787-0. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31988329

[12] N. Ebrahim, K. Shakirova, and E. Dashinimaev, "Pdx1 is the cornerstone of pancreatic beta-cell functions and identity," vol. 9, p. 1091757. doi: 10.3389/fmolb.2022.1091757. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/36589234

[13] Y. Jiang, C. Chen, L. N. Randolph, S. Ye, X. Zhang, X. Bao, and X. L. Lian, "Generation of pancreatic progenitors from human pluripotent stem cells by small molecules," vol. 16, no. 9, pp. 2395–2409. doi: 10.1016/j.stemcr.2021.07.021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/34450037

[14] Sartorius, "Ambr® 15 - advanced microbioreactor system." [Online]. Available: https://www.sartorius.com/en/products/fermentation-bioreactors/ambr-multi-parallel-bioreactors/ambr-15-cell-culture

[15] G. Alessandra, M. Algerta, M. Paola, S. Carsten, L. Cristina, M. Paolo, M. Elisa, T. Gabriella, and P. Carla, "Shaping pancreatic $\beta$-cell differentiation and functioning: The influence of mechanotransduction," vol. 9, no. 2, p. 413. [Online]. Available: https://www.mdpi.com/2073-4409/9/2/413

[16] R. Myers and D. Montgomery, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 3rd ed. Wiley. ISBN 9780471581000

[17] C. Audet, S. Le Digabel, V. R. Montplaisir, and C. Tribes, "Algorithm 1027: Nomad version 4: Nonlinear optimization with the mads algorithm," vol. 48, no. 3, pp. 1–22. doi: 10.1145/3544489

[18] B. K. Gage, T. D. Webber, and T. J. Kieffer, "Initial cell seeding density influences pancreatic endocrine development during in vitro differentiation of human embryonic stem cells," vol. 8, no. 12, p. e82076. doi: 10.1371/journal.pone.0082076. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24324748

[19] S. Takizawa-Shirasawa, S. Yoshie, F. Yue, A. Mogi, T. Yokoyama, D. Tomotsune, and K. Sasaki, "Fgf7 and cell density are required for final differentiation of pancreatic amylase-positive cells from human es cells," vol. 354, no. 3, pp. 751–9. doi: 10.1007/s00441-013-1695-6. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23996199

[20] T. Toyoda, S. Mae, H. Tanaka, Y. Kondo, M. Funato, Y. Hosokawa, T. Sudo, Y. Kawaguchi, and K. Osafune, "Cell aggregation optimizes the differentiation of human escs and ipscs into pancreatic bud-like progenitor cells," vol. 14, no. 2, pp. 185–97. doi: 10.1016/j.scr.2015.01.007. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25665923

[21] K. A. D'Amour, A. D. Agulnick, S. Eliazer, O. G. Kelly, E. Kroon, and E. E. Baetge, "Efficient differentiation of human embryonic stem cells to definitive endoderm," vol. 23, no. 12, pp. 1534–41. doi: 10.1038/nbt1163. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/16258519

[22] X. Xu, V. L. Browning, and J. S. Odorico, "Activin, bmp and fgf pathways cooperate to promote endoderm and pancreatic lineage cell differentiation from human embryonic

stem cells," vol. 128, no. 7-10, pp. 412–27. doi: 10.1016/j.mod.2011.08.001. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21855631

[23] S. Ghorbani-Dalini, N. Azarpira, M. H. Sangtarash, H. R. Soleimanpour-Lichaei, R. Yaghobi, S. Lorzadeh, A. Sabet, M. Sarshar, and I. H. Al-Abdullah, "Optimization of activin-a: a breakthrough in differentiation of human induced pluripotent stem cell into definitive endoderm," vol. 10, no. 5, p. 215. doi: 10.1007/s13205-020-02215-3. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/32355589

[24] C. H. Cho, N. R. Hannan, F. M. Docherty, H. M. Docherty, M. Joao Lima, M. W. Trotter, K. Docherty, and L. Vallier, "Inhibition of activin/nodal signalling is necessary for pancreatic differentiation of human pluripotent stem cells," vol. 55, no. 12, pp. 3284–95. doi: 10.1007/s00125-012-2687-x. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23011350

[25] S. J. Micallef, M. E. Janes, K. Knezevic, R. P. Davis, A. G. Elefanty, and E. G. Stanley, "Retinoic acid induces pdx1-positive endoderm in differentiating mouse embryonic stem cells," vol. 54, no. 2, pp. 301–5. doi: 10.2337/diabetes.54.2.301. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/15585742

[26] D. S. Lorberbaum, S. Kishore, C. Rosselot, D. Sarbaugh, E. P. Brooks, E. Aragon, S. Xuan, O. Simon, D. Ghosh, C. Mendelsohn, P. Gadue, and L. Sussel, "Retinoic acid signaling within pancreatic endocrine progenitors regulates mouse and human beta cell specification," vol. 147, no. 12. doi: 10.1242/dev.189977. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/32467243

[27] O. E. Olaniru, U. Kadolsky, S. Kannambath, H. Vaikkinen, K. Fung, P. Dhami, and S. J. Persaud, "Single-cell transcriptomic and spatial landscapes of the developing

human pancreas," vol. 35, no. 1, pp. 184–199 e5. doi: 10.1016/j.cmet.2022.11.009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/36513063

[28] J. S. Alford, "Bioprocess control: Advances and challenges," vol. 30, no. 10-12, pp. 1464–1475. doi: 10.1016/j.compchemeng.2006.05.039

[29] A. S. Rathore, S. Mishra, S. Nikita, and P. Priyanka, "Bioprocess control: Current progress and future perspectives," vol. 11, no. 6. doi: 10.3390/life11060557. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/34199245

[30] P. P. Mondal, A. Galodha, V. K. Verma, V. Singh, P. L. Show, M. K. Awasthi, B. Lall, S. Anees, K. Pollmann, and R. Jain, "Review on machine learning-based bioprocess optimization, monitoring, and control systems," vol. 370, p. 128523. doi: 10.1016/j.biortech.2022.128523. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/36565820

[31] C. Li, P. Zheng, Y. Yin, B. Wang, and L. Wang, "Deep reinforcement learning in smart manufacturing: A review and prospects," vol. 40, pp. 75–101. doi: 10.1016/j.cirpj.2022.11.003

[32] S. Kumar, T. Gopi, N. Harikeerthana, M. K. Gupta, V. Gaur, G. M. Krolczyk, and C. Wu, "Machine learning techniques in additive manufacturing: a state of the art review on design, processes and production control," vol. 34, no. 1, pp. 21–55. doi: 10.1007/s10845-022-02029-5

[33] Z. Huang, M. Fey, C. Liu, E. Beysel, X. Xu, and C. Brecher, "Hybrid learning-based digital twin for manufacturing process: Modeling framework and implementation," vol. 82. doi: 10.1016/j.rcim.2023.102545

[34] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri, "Nih image to imagej: 25 years of

image analysis," vol. 9, no. 7, pp. 671–5. doi: 10.1038/nmeth.2089. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22930834

[35] D. R. Stirling, M. J. Swain-Bowden, A. M. Lucas, A. E. Carpenter, B. A. Cimini, and A. Goodman, "Cellprofiler 4: improvements in speed, utility and usability," vol. 22, no. 1, p. 433. doi: 10.1186/s12859-021-04344-9. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/34507520

[36] F. J. Leong, M. Brady, and J. O. McGee, "Correction of uneven illumination (vignetting) in digital microscopy images," vol. 56, no. 8, pp. 619–21. doi: 10.1136/jcp.56.8.619

[37] MathWorks, "Image processing toolbox™ reference," MathWorks Inc., Report. [Online]. Available: https://www.mathworks.com/help/pdf_doc/images/images_ref.pdf

[38] H. G. Barrow and J. M. Tenenbaum, "Interpreting line drawings as three-dimensional surfaces," vol. 17, no. 1, pp. 75–116. doi: 10.1016/0004-3702(81)90021-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0004370281900217

[39] R. van den Boomgaard and R. van Balen, "Methods for fast morphological image transforms using bitmapped binary images," vol. 54, no. 3, pp. 252–258. doi: 10.1016/1049-9652(92)90055-3. [Online]. Available: https://www.sciencedirect.com/science/article/pii/1049965292900553

[40] M. R. Peres, "Laboratory imaging and photography : best practices for photomicrography and more." [Online]. Available: http://proquest.safaribooksonline.com/?fpi=9781317593003

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," vol. 60, no. 6, pp. 84–90. doi: 10.1145/3065386

[42] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," pp. 2980–2988.

[43] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," p. arXiv:1804.02767. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2018arXiv180402767R

[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition."

[45] J. Ansel, E. Yang, H. He, N. Gimelshein, A. Jain, M. Voznesensky, B. Bao, P. Bell, D. Berard, E. Burovski, G. Chauhan, A. Chourdia, W. Constable, A. Desmaison, Z. DeVito, E. Ellison, W. Feng, J. Gong, M. Gschwind, B. Hirsh, S. Huang, K. Kalambarkar, L. Kirsch, M. Lazos, M. Lezcano, Y. Liang, J. Liang, Y. Lu, C. K. Luk, B. Maher, Y. Pan, C. Puhrsch, M. Reso, M. Saroufim, M. Y. Siraichi, H. Suk, S. Zhang, M. Suo, P. Tillet, X. Zhao, E. Wang, K. Zhou, R. Zou, X. Wang, A. Mathews, W. Wen, G. Chanan, P. Wu, and S. Chintala, "Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation," in *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*. Association for Computing Machinery, Conference Proceedings. doi: 10.1145/3620665.3640366 pp. 929–947–929–947.

[46] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with numpy," vol. 585, no. 7825, pp. 357–362–357–362.

[47] A. Clark, "Pillow (pil fork) documentation." [Online]. Available: https://buildmedia.readthedocs.org/media/pdf/pillow/latest/pillow.pdf

[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization."

[49] T. Schaul, I. Antonoglou, and D. Silver, "Unit tests for stochastic optimization."

[50] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. Contributors, "Scipy 1.0: Fundamental algorithms for scientific computing in python," vol. 17, pp. 261–272–261–272.

[51] M. Mohammadi and S. Das, "Snn: Stacked neural networks," p. arXiv:1605.08512. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2016arXiv160508512M

[52] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press. ISBN 9780262035613. [Online]. Available: https://books.google.ca/books?id=Np9SDQAAQBAJ

[53] R. Caruana, "Multitask learning," vol. 28, no. 1, pp. 41–75. doi: 10.1023/A:1007379606734

[54] D. Bertsimas and N. Kallus, "From predictive to prescriptive analytics." doi: 10.48550/arXiv.1402.5481