National College of Ireland

Machine learning to forecast cell growth in bioreactor using Raman spectroscopy

MSc Research Project

MSc in Data Analytics

Vikas Khatri

Student ID: x21164894

School of Computing

National College of Ireland

Supervisor: Sasirekha Palaniswamy

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Vikas Khatri		
Student ID:	x21164894		
Programme:	MSc in Data Analytics	**Year:**	2023
Module:	Research Project		
Supervisor:	Sasirekha Palaniswamy		
Submission Due Date:	14 Dec 2023		
Project Title:	Machine Learning to forecast Cell growth in Bioreactor using Raman spectroscopy		
Word Count:	5425	**Page Count**	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:

Date: 14 Dec 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Machine learning to forecast cell growth in bioreactor using Raman spectroscopy.

Vikas Khatri

x21164894

Abstract

Practical challenge within a biopharmaceutical organization is addressed in this research that specializes in manufacturing a product, crucial for treating life-threatening diseases. The research draws data from their research laboratory and aims to predict cell growth in bioreactors by leveraging information from both the bioreactors and Raman spectroscopy. Unlike many existing studies that solely rely on data from bioreactors or Raman spectroscopy, our approach involves combining datasets. This comprehensive dataset will offer the organization a real-time overview and enhanced understanding of bioreactor process parameters. Traditional model such as partial linear regression (PLS) is built using SIMCA (Soft independent modelling by class analogy) software as part of this research. RMSEE and RMSECV performance indicators are used to evaluate the performance of the models to predict the cell growth. Predicting cell growth will contribute to achieving a batch on the initial attempt, thereby reducing production costs, maintaining supply chain demand, minimizing waste, decreasing labour hours, and lowering utility expenses.

1 Introduction

Bioreactors play a crucial role in influencing cell growth, differentiation and tissue creation by suppling nutrients and biomimetic stimuli under regulated conditions. The main advantage of bioreactors lies in their capability to cultivate a significant quantity of cells over an extended duration within a consistent environment. Cells that have undergone cultivation are employed in the production of biologics, including vaccines, therapeutic proteins, antibodies, and cell therapy products. This process has been extensively applied to enhance the proliferation of various cell types such as mesenchymal stem cells, induced pluripotent stem cells, CAR T cells, and red blood cells. (Stephenson and Grayson, 2018). In the realm of cell therapy, the substitution of impaired cells with fresh, healthy cells occurs, aiming to regulate the function of the patient's cells either by influencing gene expression, direct interaction, or eliminating disease-causing or malfunctioning cells employing immune cells. (AstraZeneca, 2023). In 2018, the combined revenue from the top five best-selling recombinant proteins exceeded US$48 billion. The compound annual growth rate (CAGR) for antibody revenue experienced a remarkable 20% surge from 2004 to 2014, although sustaining such growth rates becomes challenging as the overall market size expands, as reported by BioProcess International (2023).

Cultivating cells in a manufacturing setting is a time-consuming and costly process, extending over weeks or months. Given the high cost associated with cell growth, treatments utilizing

these cells become prohibitively expensive for many patients. To enhance accessibility for a larger patient population, it is crucial to reduce the overall cost of treatments.

This study aims to forecast cell growth. The significance of bioreactor control has increased due to challenges in managing the behaviour of living cells within cell culture systems. Employing predictive models offers the advantage of a data-driven approach, expediting the process development (International BioPharma, 2022). This will enable organizations to anticipate any outlier during batch runs and decide on corrective actions quickly in the event of failure. Ensuring the accuracy of the batch from the outset will result in lower production costs, sustained supply chain demand, waste reduction, decreased labour hours, and reduced utility expenses. This will in turn lead to a reduction in product prices, benefiting both the organization and the patients by ensuring more affordable treatment options. Furthermore, the model will enable the organization to monitor batches in real-time using SIMCA. Any deviation in process parameters will trigger real-time alarms or warnings, providing the laboratory with timely information to intervene.

The objective is to predict cell growth within the bioreactor, utilizing data from both the bioreactor and Raman spectroscopy. While previous research has often relied on either Raman spectroscopy or bioreactor data alone, this study integrates parameters from both sources. These parameters include the bioreactor's temperature, temperature setpoint, pO2 value, pO2 setpoint, pH setpoint, and pH value of the media in the bioreactor. A comprehensive list of all parameters is provided in the Appendix. The Raman spectroscopy method measures the dependent variable, viable cell density (VCD). The organization follows specific rules for assessing cell growth determined by the measured VCD values. A conventional model using partial least squares regression is developed based on insights from the literature review. Performance indicators, namely RMSEE and RMSEECV, will be employed to assess the models' effectiveness in predicting cell growth.

2 Related Work

2.1 Understanding of cell growth

Anane, Knudsen and Wilson (2021) present a well-structured study on the impact of dissolved oxygen (dO2) gradients on CHO cell cultures in a multi-compartment scale-down simulator. The inclusion of a plug-flow reactor (PFR) is notable, as it addresses the fragility of mammalian cells, which has limited the application of such simulators in mammalian cell cultivation processes. The findings indicate a clear switch in CHO cell metabolism in response to dO2 gradients beyond a residence time threshold of 90s in the PFR. The study observes an accumulation of lactate and a decline in viable cell density, impacting the product quality. Interestingly, recombinant protein productivity remains unaffected, emphasizing the complexity of cellular responses in such environments. Clarity on the relevance of the study in

the broader context of bioprocess development and its potential impact on industrial applications could be highlighted in the conclusion.

Carpio (2020) highlights the evolution of mammalian cell culture scale-up over the past decade and the challenges associated with it. The focus is on the significant advancements in process development, particularly the shift from microplates and shake flasks to high-throughput automated micro bioreactors and mini bioreactors. These innovations have not only accelerated the selection of optimal clones but also doubled average product titers. The importance of scaling tools and analytical methods for seamless scale-up is discussed, offering potential solutions to improve efficiency and reduce costs.

Using Raman spectral characteristics and a machine-learning model, Yamamoto *et al*. (2023) describe a unique method for predicting the growth/no growth response of unknown bacteria. 21 strains are isolated from seven fresh-cut vegetables in the study, and each strain's Raman spectra and growth/no growth responses are recorded. With 90% accuracy, the created artificial neural network (ANN) model forecasts the growth/no growth of 21 unidentified microorganisms. This research emphasises the potential of Raman spectroscopy and machine learning for the identification of unknown bacterial growth.

Aryani *et al.* (2015) investigates the impact of strain variability on microbial growth kinetics using twenty Listeria monocytogenes strains. The maximum specific growth rate was assessed in relation to pH, water activity concentration [NaCl], undissociated lactic acid concentration ([HA]), and temperature (T). The study employs secondary growth models to estimate cardinal growth parameters for pH, [NaCl], [HA], and T. The results emphasize the importance of understanding strain variability for realistic predictions of microbial growth kinetics in food products. Barbosa *et al.* (1994) shed information on temperature-dependent differences in different serotypes of Listeria, which provides important insights into the growth pattern associated with these strains. The results are essential for comprehending how Listeria behaves under food processing and storage circumstances.

Coroller, *et al.* (2012) The model integrates growth, growth/no-growth boundaries, and inactivation phases, considering factors like temperature, pH, water activity, lactic acid, and sorbic acid. Data from diverse sources were utilized for model development and validation. The proposed model demonstrates high accuracy (62%-87% correct predictions) and low median errors (<0.34 log10(CFU/mL)).

Dengremont and Membré (1995) employ a predictive microbiological approach to quantify Staphylococcus aureus growth in food, considering factors like temperature, pH, and NaCl concentration. The linear and nonlinear models are compared. The nonlinear model, addressing the complex interactions, outperforms the linear one, offering better fitness and parameter interpretability. This paper is well-structured, methodologically sound, and makes a significant contribution to predictive microbiology.

2.2 A critical survey of techniques in predicting cell growth

Wold, Sjöström and Eriksson (2001) provide a comprehensive review of Partial Least Squares Regression (PLSR) as a widely used multivariate analysis tool in chemistry and technology. PLSR, a two-block predictive PLS model, facilitates the relationship modelling between two data matrices, X and Y, handling collinear, noisy, and incomplete variables. It is discussed in this paper that underlying model, assumptions, and diagnostics, emphasizing PLSR's ability to improve precision with an increasing number of relevant variables and observations.

Tulsyan, Garvin and Ündey (2018) propose a novel solution using machine-learning methods to generate in silico batches, mitigating the need for historical data. The approach combines hardware exploitation and algorithm development, demonstrating efficacy through case studies. In contrast to traditional multivariate monitoring, the proposed method detects weak signals in real-time, reducing the risk of overlooking critical process deviations using SIMCA software.

Gibbons *et al*. (2021) explore Process Analytical Technology (PAT) tools, specifically Raman spectroscopy and chemometric modelling, the research develops Partial Least Squares (PLS) regression models for real-time monitoring of glycation and glycosylation profiles. While effective at a small scale, challenges arise at manufacturing scale, indicating the importance of scale considerations in model development. To enhance model robustness, the study incorporates manufacturing scale data, significantly improving predictions, particularly for glycosylation.

Banner *et al*. (2021) conducted a review spanning the decade from 2010 to 2020 to assess the application of data analytics in the biopharmaceutical industry. Their findings highlighted a prevailing trend toward the utilization of machine learning algorithms within this sector.
It highlights a shift from traditional multivariate data analysis to a broader use of machine learning (ML) algorithms, driven by advancements in bioreactor technologies and the adoption of process analytical technology. The study reviews prominent algorithms, such as partial least squares (PLS) and neural networks (NN), applied to diverse datasets. The analysis emphasizes the prevalent use of PLS, especially in PAT applications, and anticipates continued integration of ML for improved process understanding and optimization in biomanufacturing. Alavijeh *et al.* (2022) explores the evolving landscape of bioreactor scale-up in the pharmaceutical industry, emphasizing the shift from traditional, rule-of-thumb methods to digital strategies. Focusing on the challenges of scaling biological processes, the paper discusses the limitations of existing approaches and delves into the potential of digital tools, including knowledge-driven and data-driven techniques.

Rafieyan *et al*. (2023) explore predicting cell behaviour on cardiac tissue engineering (CTE) using machine learning (ML). A novel software, MLATE, was developed to predict cell behaviour on CTE scaffolds based on materials, cell lines, and fabrication methods. ML models

demonstrated substantial predictive power, with ensemble techniques achieving 93% accuracy. MLATE's efficiency and short prediction run times further enhance its utility.

Rashedi *et al.* (2022) addresses the challenges that arise in biopharmaceutical process control by proposing a Model Predictive Controller (MPC) based on a linear machine learning model. The bioprocess aims to maximize cell growth and metabolite production and faces complexities such as limited measurements and process variations. The linear MPC is developed using a hybrid model combining machine learning and glucose mass balance equations, demonstrated a 2% improvement in final protein production compared to traditional methods.

A comparative assessment of various techniques for predicting cell growth in bioreactors was conducted, examining methods employed in related studies. Traditional approaches, like regression models, were initially employed, but due to their limitations surrounding diverse cell type datasets, a global model was subsequently developed using deep learning models. Notably, some studies either focused solely on bioreactors or on the Raman system leading to suboptimal model performance. This highlights the need to incorporate both systems (Bioreactor and Raman) to ensure accurate cell growth (VCD) prediction. SIMCA tool is used to build the model and predict the cell growth (VCD). PLS algorithm is used and the performance of these models was assessed using common metrics such as RMSEE and RMSEECV.

3 Research Methodology

The section research methodology is a systematic and structured approach used in the experiment of cell growth in bioreactor using Raman spectroscopy. SIMCA software is used to make the model to generate precise and accurate results as the tool is used in the organisation to build the models.

3.1 Ethical Concerns

The chosen dataset is taken from organisation laboratory and does not include any personally identifiable information related to individuals. Also, the dataset does not include information about race, religion, and sexual orientation. The dataset only includes data about the equipment therefore GDPR does not apply.

3.2 Data Collection

Company data is used in this study. The data is generated in the lab from bioreactors which is controlled by DCU (digital control unit). DCU is connected with Modular Fermentation and Culture System (MFCS) which analyse/process the bioreactor data, enable trending and

provide recipe creation which is widely used in the lab. MFCS is connected to site historian/PI (Process Intelligence) server to store and provide data to different systems. MFCS sits on the lab network and PI system sits on enterprise network and these are segregated by firewall. PI system makes bioreactor data available on enterprise network, which is used by SIMCA online. SIMCA online collects the batch data of bioreactor's parameters and make the data available to SIMCA client which sits in the lab.

Additionally, Raman spectroscopy is a standalone equipment in the lab, the probes of the Raman are directly connected with bioreactor to measure the VCD. For each batch, VCD is measured at a regular interval. The VCD data is exported as comma-separated values (CSV) file from Raman system and imported to SIMCA in the lab.

The system architecture of all the systems stated above is shown below. It represents the way equipment are connected with each other in the lab and with organisation enterprise network, where the data is being generated and how the data is being transferred to different systems.

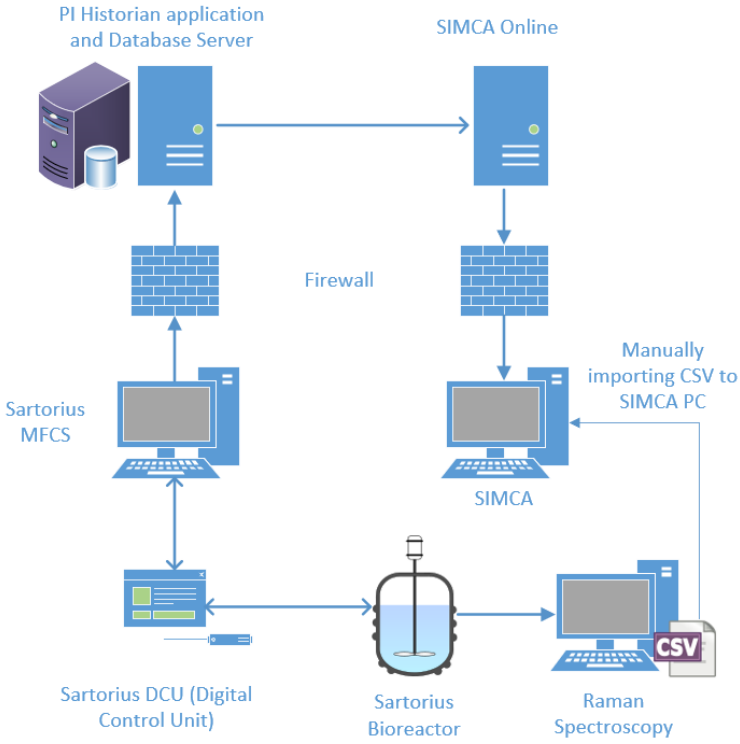

Figure 1: System Architecture

Typically, the data are measured throughout time at regular intervals, such as once a day, once an hour, or once a minute. For a number of variables, these intervals frequently vary.

3.2.1 Biostat B-DCU Bioreactor and MFCS

The Biostat B-DCU (Digital Control Unit) is a specialized bioreactor tailored to meet the needs of process optimization and characterization in the biotech and biopharmaceutical sector. With

advanced functionality and an unparalleled range of options for cell culture and microbial processes, it serves as the optimal scale-down model for large-scale processes Sartorious (2023a).

Figure 2: Bioreactors with DCU tower & MFCS graphical user interface Sartorious (2023a) Sartorious (2023b)

MFCS introduces a new standard in bioprocess data management and automation. With dependable data acquisition, effective trend monitoring, and advanced recipe control, it proves to be an ideal tool for both upstream and downstream processes, regardless of whether single-use or reusable systems are favored. In R&D environments, MFCS stands as the solution for establishing robust and reproducible processes Sartorious (2023b).

3.2.2 Raman Rxn2 analyzer

Effectively leverage the capabilities of Raman spectroscopy with the Raman Rxn2 analyser. Tailored for use in analytical laboratories with model transfer functionalities, the Raman Rxn2 is a trusted tool for routine sample identification, R&D project support, early process development, and in situ analysis during scale-up processes. Whether in a benchtop configuration or on a mobile wheeled cart, the Raman Rxn2 offers flexibility and portability in process development laboratories. The convenience of a single base unit that accommodates up to four probes, coupled with an intuitive embedded control software accessible via touchscreen or remote interface, ensures reliable real-time in situ measurements Endress + Hauser (2023). The analyser's sequential operation enables swift analysis per channel, and programmable channel interrogation transforms acquired Raman spectra into valuable process knowledge through integrated multivariate predictors. The Raman Rxn2 is particularly well-suited for applications in bioprocess monitoring and control, cell culture, fermentation, and downstream operations Endress + Hauser (2022).

Figure 3: Raman spectroscopy – RXn2 Analyzer Endress + Hauser (2022) Endress + Hauser (2023)

3.2.3 Aveva PI

AVEVA (2023) PI (Process Intelligence) System, formerly known as OSIsoft PI System, stands out as a leading data management solution designed specifically to address the challenges prevalent in industrial environments. This innovative system enables the collection and storage of data from diverse locations and sources, extracting valuable insights rapidly to optimize business processes—all within a flexible, no-code environment. Offering deeper operational insights, swift analysis of crucial data, and enhanced visibility of remote assets and IIoT (Industrial Internet of Things) sensors, AVEVA PI System contributes to more efficient and sustainable operations.

It excels in collecting real-time data from various assets, including legacy, proprietary, remote, mobile, and IIoT devices. AVEVA PI System seamlessly connects organizations to their data, irrespective of location or format, and has the capacity to store decades' worth of data with sub-second granularity.

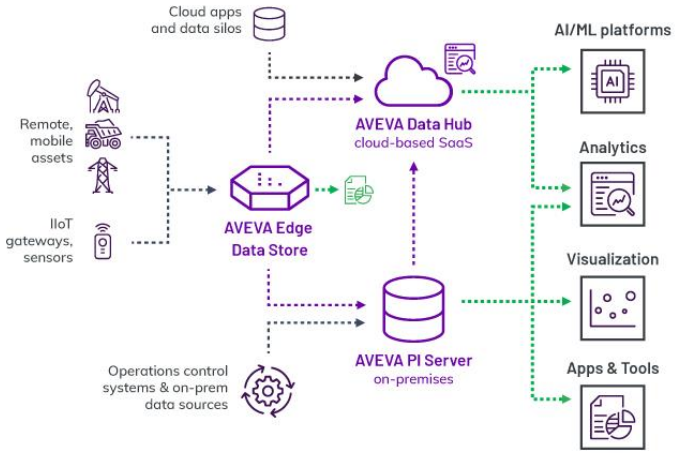

Figure 4: PI system flow and architecture AVEVA (2023)

3.2.4 SIMCA

SIMCA (2022) facilitates real-time multivariate process monitoring and control by leveraging SIMCA models and data sourced from systems like process historians. SIMCA is a versatile software designed for comprehensive multivariate data analysis, capable of handling any data that can be converted into numerical form. Executed on a desktop computer, SIMCA conducts data analysis using models stored in SIMCA project files (.usp files).

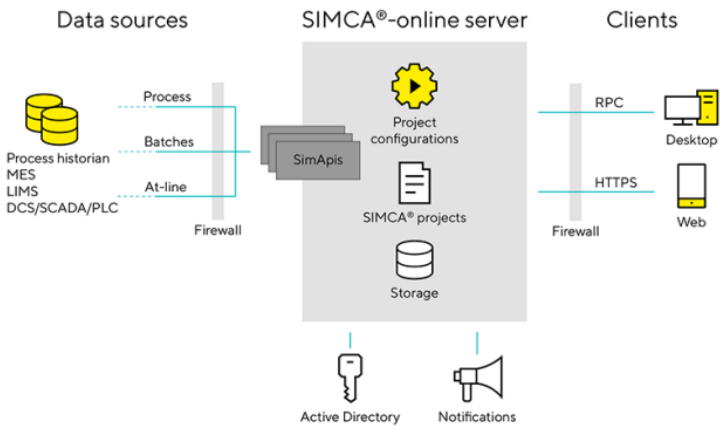

Figure 5: SIMCA architecture SIMCA (2022)

3.2.5 Process Parameters

See Appendix for list of parameters.

3.3 Data processing

In this research, comprehensive data processing was performed using SIMCA software to prepare the data to build the PLS model. There are 26 batches used to build the model, each batch has 36 process parameters which are generated from bioreactors and imported to SIMCA. The process of producing several identical products all at once is known as batch production. Every batch passes through each stage of the manufacturing process in tandem with the others. For each batch, the value of each parameter is observed every 3 minutes. Since the batches were produced during different dates and months, SIMCA created a new parameter called 'Time days', which represents the stage of the batch in a day e.g., 1st day, 4th day, etc, rather than on a specific time or day, so all the batches can be compared as time days. SIMCA did this in order to process the data and build the model. The below figure shows the "Time days" parameter for each batch. Each colour on the trend represents batch, whereas the x axis represents primary id (for each timestamp there is a unique primary id) and the y axis shows the number of days the batch ran for.

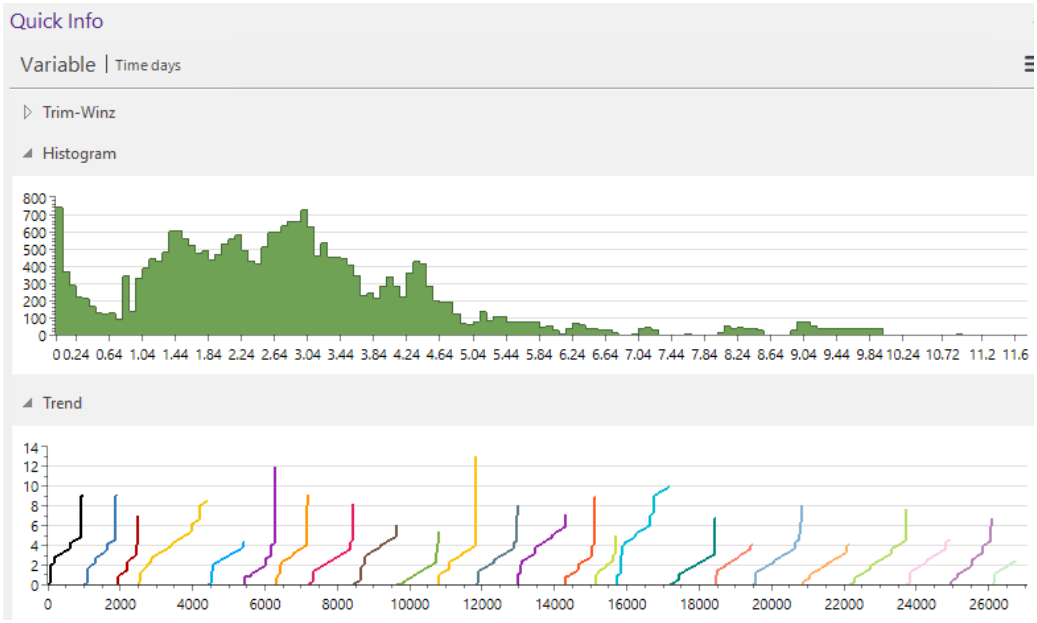

Figure 6: Time days parameter representing each batch's timeline

Further on, highly corelated parameters such as SUBS A ST PT%, SUBS B ST PT%, BASESUB st pt_value difference, SUBSA st pt_value difference and SUBSB st pt_value difference are removed from this research.

Additionally, out of spec data (Outliers) were removed from each parameter (if any) in consultation with lab scientist so that the correct value to be fed to model. All the null values are auto fill by SIMCA software.

The below figure shows the trend for AIRSP Value parameter, each colour represents the value of parameter for each batch. As part of the data cleaning, values greater than 1.2 and less than 0.8 were removed as these were the out of spec values. This is one of the examples of removing out of spec values from parameters.

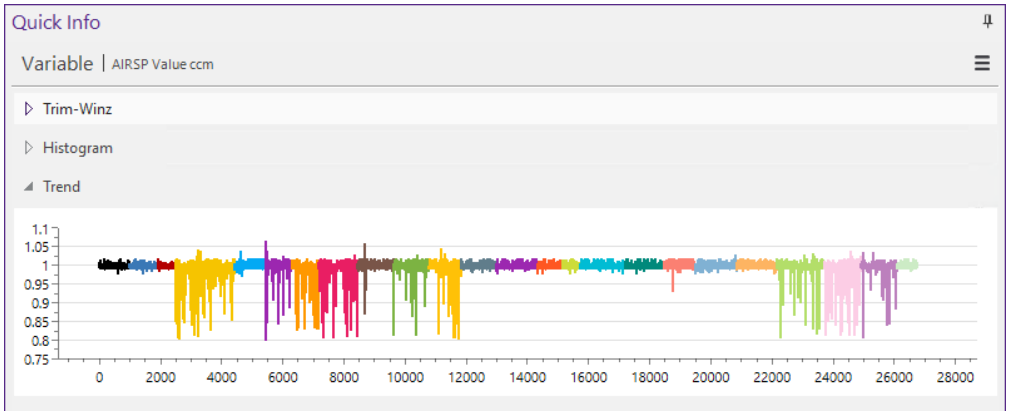

Figure 7: AIRSP Value for each batch

Below figure shows the pH value for each batch. On the left-hand side, all the batches are listed with a colour code, the graph shows the pH value for those batches. X axis shows the "Time days" parameter and Y axis shows the pH value.

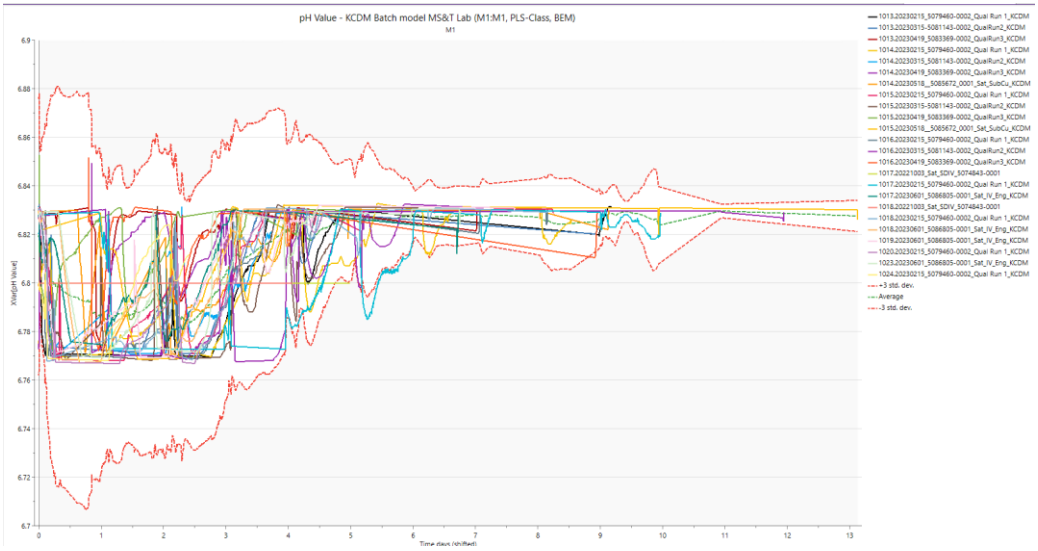

Figure 8: pH value for each batch

Similarly let's have a look at the JTEMP Value parameters for each batch in below figure.

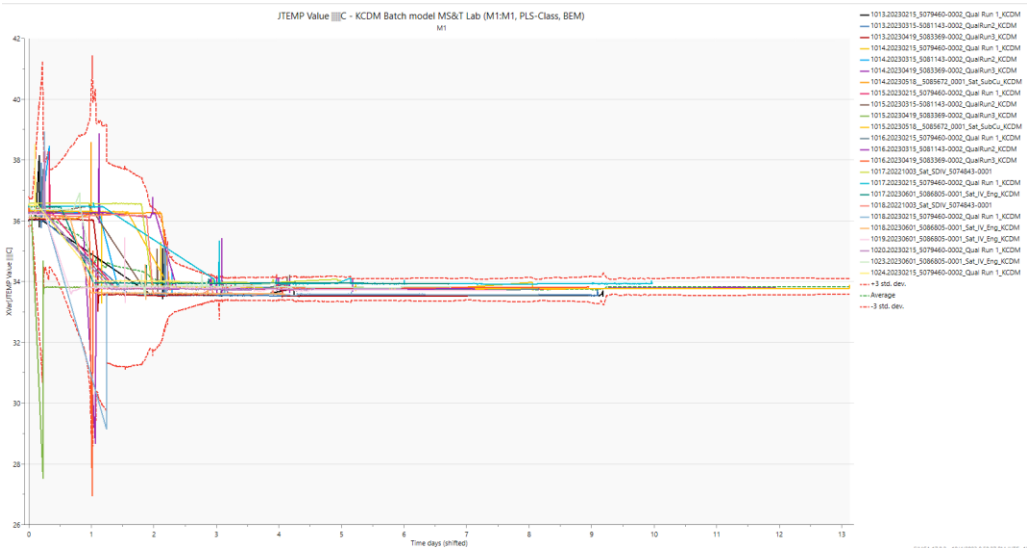

Figure 9: JTEMP value for each batch

4 Design Specification

Figure. 10 illustrates the architecture for design specification utilized in the study and demonstrates the processes carried out in this research from the beginning to the execution. Company data is used to conduct this research. Bioreactor data and data from Raman spectroscopy are used and imported to SIMCA software. SIMCA software is used to build the model.

SIMCA automatically create the 'Time Days' parameter once data is imported to calculate batch data in time days so that parameters can be trended and analysed with respect to time in dates irrespective of the date of the batch. SIMCA software is smart enough to take care of null values. Once trends were ready, out of specification data was removed from the parameters for each batch, additionally highly correlated parameters are removed from model building.

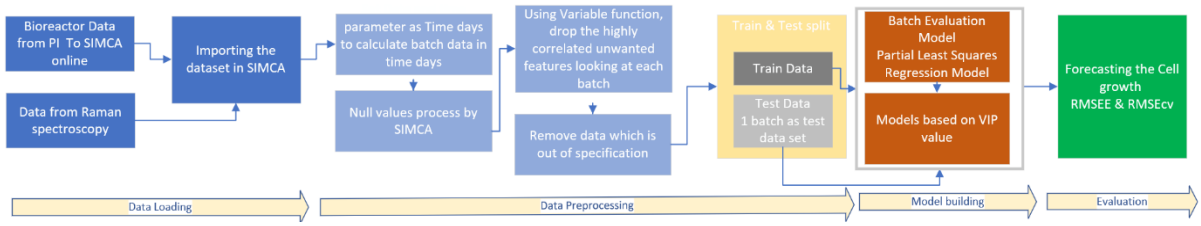

Figure 10: Design specification architecture

To create a model that can predict cell growth, a partial least squares regression model is built.

Partial least squares (PLS) regression is a technique that reduces the predictors to a smaller set of uncorrelated components and performs least squares regression on these components, instead of on the original data. When there are more predictors than data or when predictors are very collinear, PLS regression can be very helpful because, typically least-squares regression either yields coefficients with large standard errors or fails entirely. Unlike multiple regression, PLS does not presume that the predictors are fixed. As a result, PLS is more resilient to measurement uncertainty since the predictors can be measured with inaccuracy. Additionally, Partial Least Squares has the ability to model several outcome variables, which is a significant advantage. Multiple outcome variables are not something that many machine learning and statistics models can handle directly.

The VIP (Variable Importance for the projection) plot summarizes the importance of the variable both to explain X axis and to correlate to Y axis. Three different models are evaluated based on VIP value. The models are evaluated on the basis of root mean square error of estimation (RMSEE) and root mean square error of cross validation (RMSECV) value of the models.

RMSEE is computed as $\sqrt{(\sum(Yobs-Ypred)2/(N-1-A))}$, where Yobs refers to the observations which are imported to the model and Ypred refers to the prediction value of dependent variable (VCD). RMSEE measures the fit of the model.

RMSEcv – An alternative predictivity measure for the model is available through summarizing the cross-validation residuals of the observations in the dataset. The predictivity measure obtained is called RMSECV.

5 Implementation

The methods used to complete the assignment for the study are briefly discussed in this section.

5.1 Experimental Setup

Organisation's infrastructure is used to carry out this study. SIMCA workstation in the lab is used to configure the task. The lab workstation with the hardware configuration of 64-bit Windows 10 OS, Intel(R) Core (TM) i5-10500 CPU @ 3.10GHz Processor and 32GB of RAM are used. The model training is done on the local workstation using SIMCA software.

5.2 Implementation of batch level model

Batch level models use information from the process data collected during data pre-processing. It is combined with batch conditions like offline data from Raman spectroscopy. In the batch level, PLS is used to create overview of batch similarities to make predictive models for final batch quality attributes. The objective of batch level modelling is to make a model of all the batches in order to understand how VCD is influenced by the combination of batch conditions and batch evolution. This model will be based on the batch conditions, the evolution trace matrix and when applicable the properties and quality of the complete batch.

While building the model, all the processed data is converted to batch level. Further on, the process data is combined with offline data from Raman spectroscopy which has VCD value for each batch. All parameters from each batch are included in the batch level model. 25 out of 26 batches are used as train data, one remaining batch is used as test data. When batch level model is created, SIMCA automatically converts and rearrange processed data to a batch level prediction set. SIMCA generated 32307 variables/parameters out of processed data.

The VIP (Variable Importance for the Projection) plot summarises the importance of the variables both to explain x (all variables) and to correlate to Y (VCD value from Raman spectroscopy). The VIP values are calculated for each x_k by summing the squares of the PLS

loading weights, wa_k, weighted by the amount of sum squares explained in each model component. The sum of squares of all VIP's is equal to the number of terms in the model, hence the average VIP is 1.

5.3 PLS Model based on all VIP value

The X axis of VIP plot shows more than 32000 variables, it is not possible to show the names of all the variables due to space constraints.

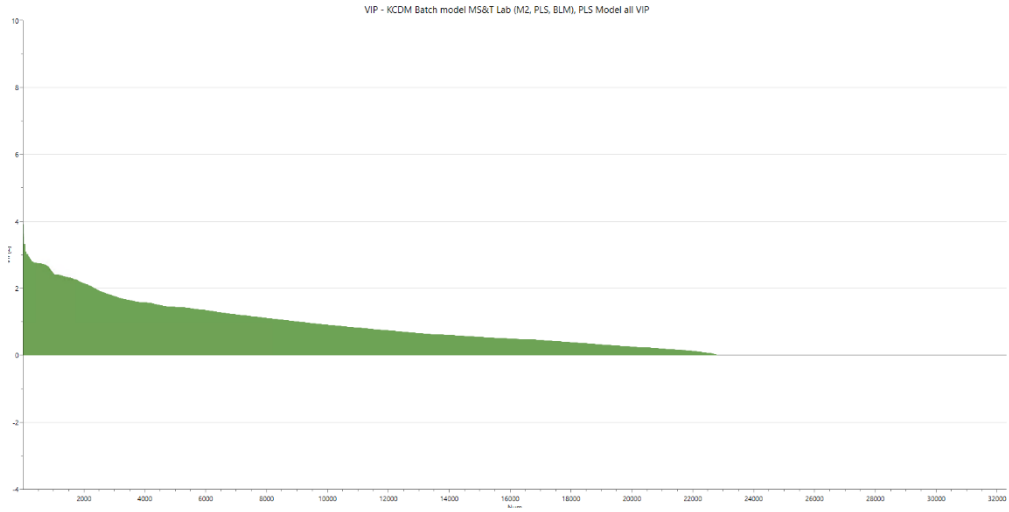

Figure 11: VIP plot

Summary of fit plot – For each model components in a PLS model, plot displays 2 bars, R2 (Green bar) which is the percent of variation of the training set. R2 is a measure fit, means how well the model fits the data. R2 close to 1 is considered as a good model. However, if training data is noisy, we can have large R2 value for a model and model can be a poor model.
Q2 (Blue bar) is the percent of variation of the training set, predicted by the model according to cross validation. Q2 indicate how well the model predicts new data, a large Q2 (Q2>0.5) indicates good predictivity. We have used two first fit for our models.
Figure 12 shows the summary of fit for model including all VIP value. R2Y cumulative is 0.962 and Q2 cumulative is 0.199. This means the model including all VIP values is a good fit model as it is close to 1. However, Q2 is way below 0.5 therefore the model does not have very good predictivity.

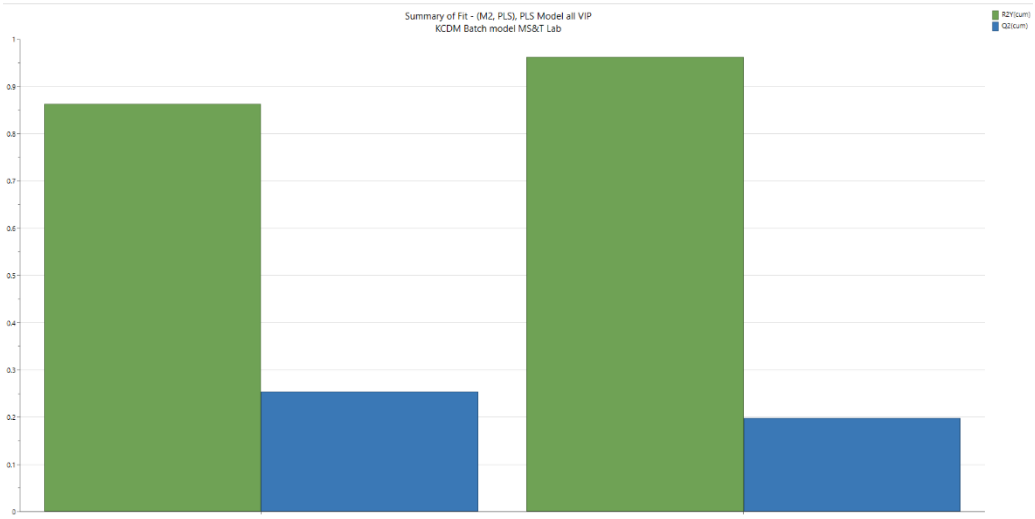

Figure 12: Summary of fit plot – all VIP value

5.4 PLS model including VIP above 1

The second model we built using VIP value above 1, that means VIP below 1 were excluded to build the model. Once done, there were 9000 variables left as these variables were above VIP 1.

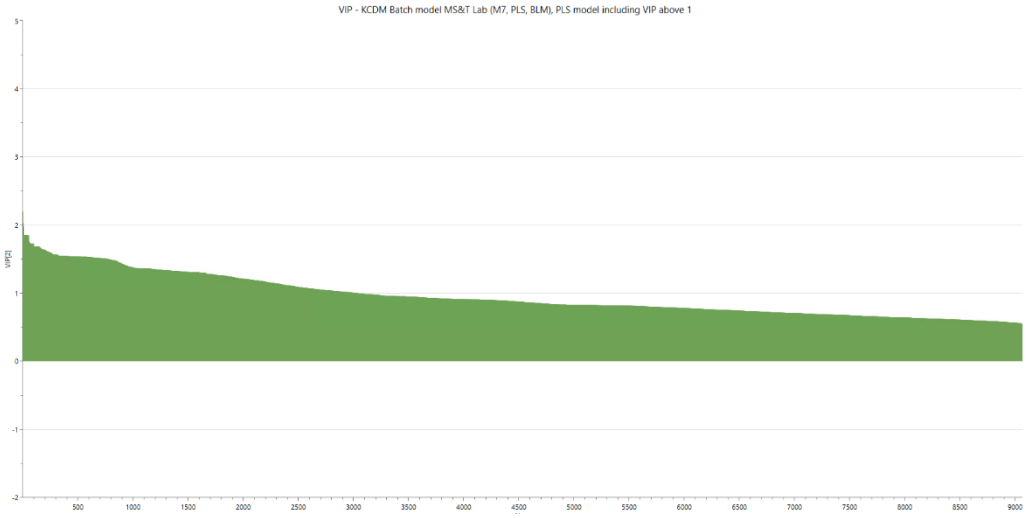

Figure 13: VIP plot – VIP above 1

Figure 14 shows the summary of fit for model including VIP value above 1. R2Y cumulative is 0.963 which is nearly same as first model and Q2 cumulative is 0.614, far more than first model and greater than the ideal value of 0.5.

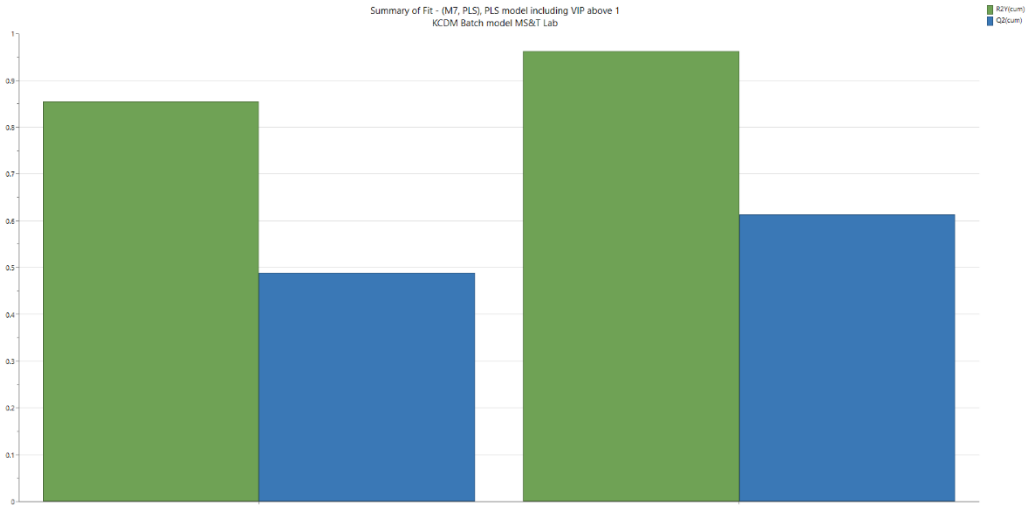

Figure 14: Summary of fit plot – VIP above 1

5.5 PLS model including VIP above 0.5

The third model we built using VIP value above 0.5, that means variables which have VIP below 0.5 were excluded to build the model. There were more than 16000 variables left as these variables were above VIP 0.5.

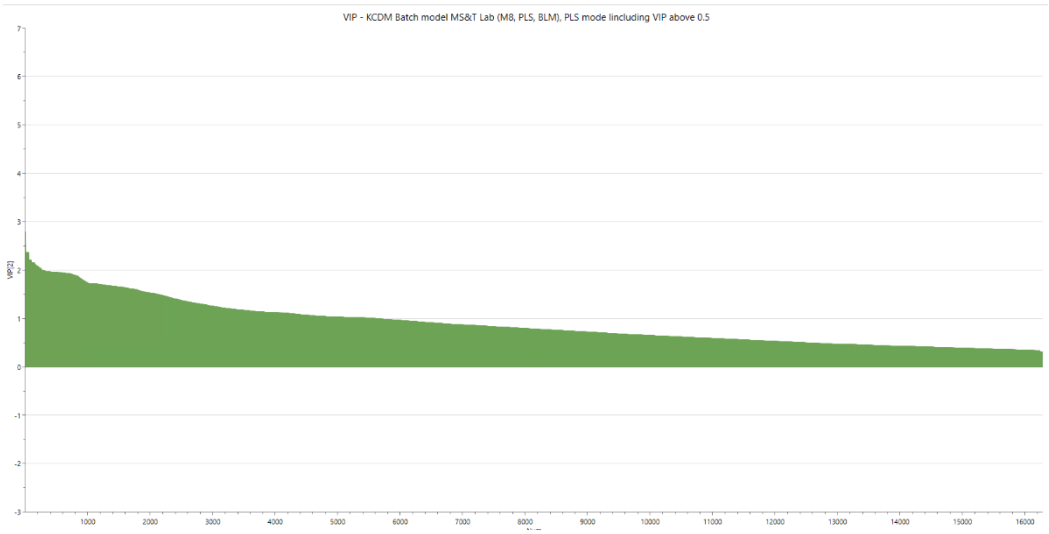

Figure 15: VIP plot – VIP above 0.5

Figure 16 shows the summary of fit for model including VIP value above 1. R2Y cumulative is 0.968 which is nearly same as first and second model and Q2 cumulative is 0.457, far more than first model and less than second model and nearby the ideal value of 0.5.

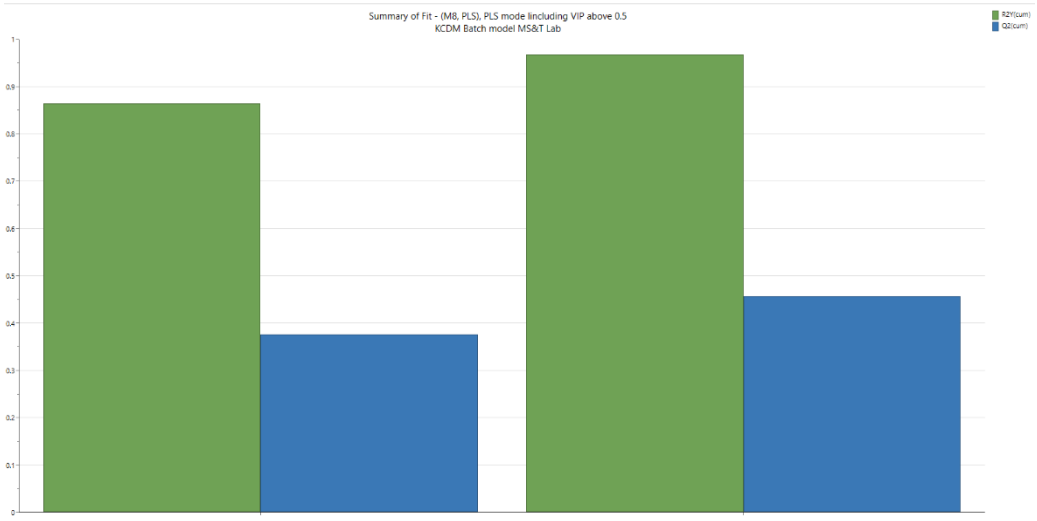

Figure 16: Summary of fit plot – VIP above 0.5

6 Evaluation

To evaluate the model, PLS model was used for all three experiments. We have shown below the observed and predicted value of VCD for each batch. Observed value (Y axis) are the one which was imported from Raman spectroscopy and model predicted (X axis) the VCD. With a good model all the points will fall close to the 45-degree line. The RMSEE in the footer indicate the fit of the observations to the model. The RMSECV is the analogous measure but estimated using the cross-validation procedure which represents the predictivity of the model.

6.1 Experiment 1 – PLS model VCD prediction based on all VIP value

Figure 17 shows the observed vs predicted plot for all VIP value. The RMSECV value is 0.171071 and RMSECV value of model is 1.19239.

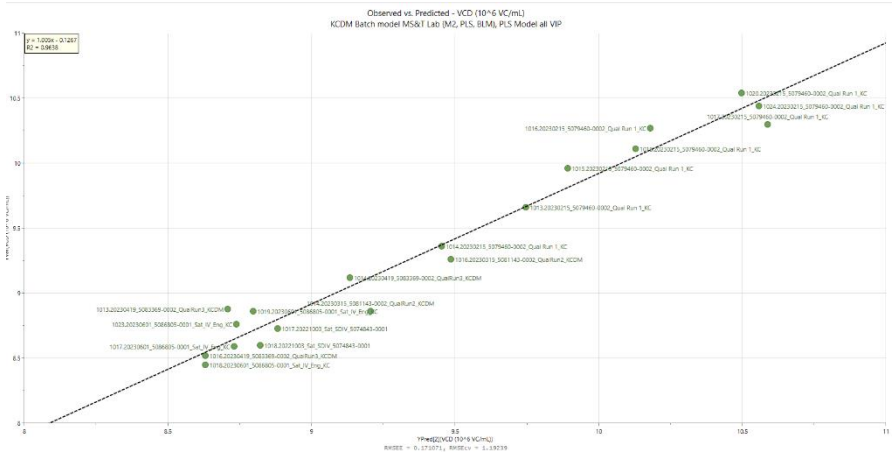

Figure 17: Observed vs predicted plot – all VIP model

6.2 Experiment 2 – PLS model VCD prediction including VIP above 1

Figure 18 shows the observed vs predicted plot for all VIP value. The RMSECV value is 0.169278 and RMSECV value of model is 1.25747.

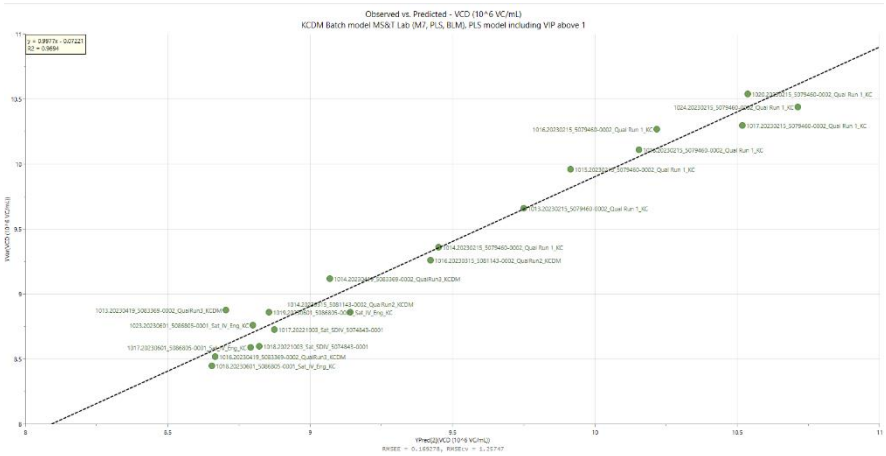

Figure 18: Observed vs predicted plot –VIP above 1 model

6.3 Experiment 3 – PLS model VCD prediction including VIP above 0.5

Figure 19 shows the observed vs predicted plot for all VIP value. The RMSECV value is 0.156985 and RMSECV value of model is 1.20109.

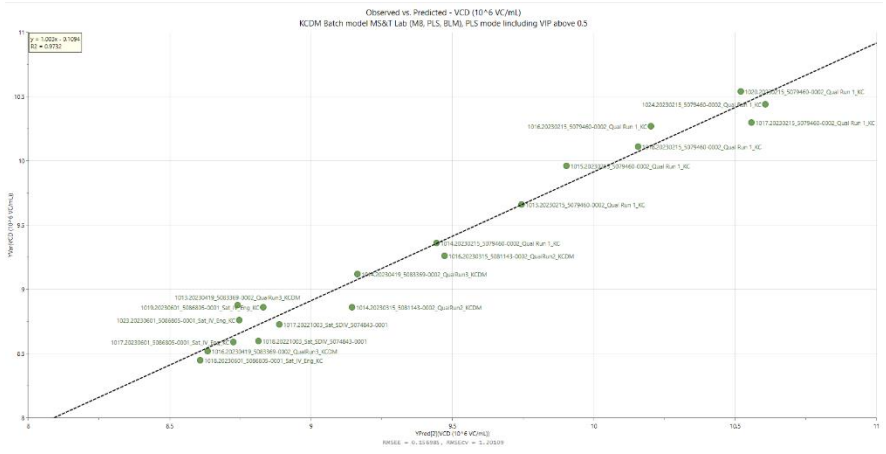

Figure 19: Observed vs predicted plot –VIP above 0.5 model

6.4 Discussion

There were a few challenges faced during the execution of this research. Batch data is transferred from MFCS system to PI system. Scientists in the lab give a unique batch number to each batch on MFCS when they start a new batch, and the process parameters are being logged into PI system so that data can be made available for different systems. When data was

first imported, the model was built using that initial data and then the model was reviewed with a lab scientist; it was found that the batch duration was not matching for several batches as those batches were run for longer days than the data was showing from PI system to SIMCA. We carried out an investigation and crossed checked the data between MFCS (Source) and PI and data matched between these two systems. Then, the data was compared between PI system and SIMCA and there were discrepancies for batch duration. After further investigation, it was found that the trigger set in SIMCA which act as batch finish signal was set up at the higher temperature. SIMCA does not have a direct connection with MFCS (Source system) and the only way SIMCA understand if the batch is finished is by setting temperature value for a duration, if batches temperature goes below that for that duration it is considered as batch is finished. The value of temperature was set higher and even without batch was finished, SIMCA interpreted that batch finished and stopped further recording the batch. This was corrected and SIMCA received full range batch data and research needed to be done again; as a result we lost a lot of time in identifying, investigating, and resolving the issue. Finally, when the issue was resolved, we were able to build the model and continue this research.

The aim of this research was to predict cell growth in the bioreactor. The selected approach was the PLS model using SIMCA software. Three experiments have been conducted and evaluated based on the model's VIP value.

Table 2 shows the results from three models. There is no big difference among the three models in the RMSEE value which represents that all three models are good fit. Similarly, the RMSEE cross validation value for three models are close enough which represents the predictivity of all three models. However, to choose best model, "PLS model with VIP above 0.5" is the best fit model and "PLS model with all VIP" has the best predictivity among three models.

Table 1: Evaluation summary

	RMSEE	RMSECV
PLS model with all VIP	0.171071	1.19239
PLS model with VIP above 1	0.169278	1.25747
PLS model with VIP above 0.5	0.156985	1.20109

Further on, we predicted the test model based on one batch data where VCD value was 9.09 and model predicted 11.88.

7 Conclusion and Future Work

This research investigated the cell growth in bioreactor at a research and investigation laboratory in a Biotechnology organisation. The research question revolves around cell prediction by predicting VCD using Raman spectroscopy. The research successfully addressed the understanding of hierarchical effective representations, building of experimental setups

using SIMCA, evaluation and result. The study clearly indicates the impact of Raman spectroscopy in predicting cell growth in bioreactors using SIMCA. Our models have shown promising results predicting VCD however result can be improved. The result has proven that "PLS model with VIP above 0.5" is the best fit model and "PLS model with all VIP" has the best predictivity amongst the three models. However, all three models have marginally differences between their results; this can be improved by including more batches to the research. The model built with SIMCA, creates real time notifications for the lab's attention if the process parameter deviates from the operational range for scientist's intervention. This model will help organisations, first to predict the cell growth and second give organisations a real time warning in the event the process deviates from standard parameters.

There was a batch limitation of only 26 batches for this research but in future, ample number of batches should be included in this research to have better models and better results. Additionally, only one batch could be included for training the model as there was a smaller number of batches to build the model. In future, more batches can be used as training models to get a precise result. SIMCA software was used to predict the model as a requirement from organisation and PLS method was used, moreover other machine learning techniques can be used to predict the model and deep learning techniques such as ANN and LSTM can be used to predict the cell growth.

References

AstraZeneca (2023) *Harnessing the power of cell therapy.* Available at: https://www.astrazeneca.com/r-d/next-generation-therapeutics/cell-therapies.html [Accessed 12 Jul 2023].

Alavijeh, M.K., Baker, I. Lee, Y.Y. and Gras, S.L (2022) 'Digitally enabled approaches for the scale up of mammalian cell bioreactors,' *Digital Chemical Engineering*, 4, p. 100040. doi: 10.1016/j.dche.2022.100040.

Anane, E., Knudsen, I.M. and Wilson, G. (2021) 'Scale-down cultivation in mammalian cell bioreactors—The effect of bioreactor mixing time on the response of CHO cells to dissolved oxygen gradients,' *Biochemical Engineering Journal*, 166, p. 107870. doi: 10.1016/j.bej.2020.107870.

Aryani, D.C., den Besten, H.M.W., Hazeleger, W.C. and Zwietering, M.H. (2015) 'Quantifying strain variability in modeling growth of Listeria monocytogenes,' *International Journal of Food Microbiology*, 208, pp. 19–29. doi: 10.1016/j.ijfoodmicro.2015.05.006.

Aveva (2023) *Aveva PI System.* Available at: PI Sys https://www.aveva.com/en/products/aveva-pi-system/tem™ | AVEVA [Accessed 07 Oct 2023]

Banner, M., Alosert, H., Spencer, C., Cheeks, M., Farid, S.S., Thomas, M. and Goldrick, S. (2021) 'A decade in review: use of data analytics within the biopharmaceutical sector', *Current Opinion in Chemical Engineering*, 34, pp. 1-10. doi: 10.1016/j.coche.2021.100758.

Barbosa, W.B., Cabedo, L., Wederguist, H.J., Sofos, J.N. and Schmidt, G.R. (1994) 'Growth variation among species and strains of listeria in culture broth', *Journal of Food Protection*, 57(9), pp. 765–769. doi: 10.4315/0362-028x-57.9.765.

BioProcess International (2023) *Supply and Demand Trends: Mammalian Biomanufacturing Industry Overview.* Available at: https://bioprocessintl.com/business/economics/forecasts-for-biomanufacturing-capacity/ [Accessed 12 Jul 2023].

Carpio, M. (2020) 'Current challenges with cell culture scale-up for biologics production', *BioPharm International*, 33(10), pp.23-27. Available at: https://www.biopharminternational.com/view/current-challenges-with-cell-culture-scale-up-for-biologics-production.

Coroller, L., Kan-King-Yu, D., Leguerinel, I., Mafart, P. and Membré J.M. (2012) 'Modelling of growth, growth/no-growth interface and nonthermal inactivation areas of Listeria in foods,'

Dengremont, E. and Membré, J. (1995) 'Statistical approach for comparison of the growth rates of five strains of Staphylococcus aureus', *Applied and Environmental Microbiology*, 61(12), pp. 4389–4395. doi: 10.1128/aem.61.12.4389-4395.1995.

Ecker, D.M. and Seymour, P. (2023) Supply and *demand trends: Mammalian biomanufacturing industry overview.* Available at: https://bioprocessintl.com/business/economics/forecasts-for-biomanufacturing-capacity/ [Accessed 14 April 2023].

Endress + Hauser. (2022) *Technical Information, Raman Rxn2*. Available at: https://bdih-download.endress.com/files/DLA/005056A500261EDE929CEF55B9EE3682/TI01608CEN_0222.pdf [Accessed 07 Oct 2023]

Endress + Hauser. (2023) Raman Rxn2 analyser. Available at: https://www.uk.endress.com/en/field-instruments-overview/optical-analysis-product-overview/raman-rxn2-analyzer?t.tabId=product-overview [Accessed 07 Oct 2023]

International Journal of Food Microbiology, 152(3), pp. 139–152. doi: 10.1016/j.ijfoodmicro.2011.09.023.

International BioPharma (2022) 'The science & business of biopharmaceuticals', *International BioPharma*, 35 (9) pp. Available at: https://cdn.sanity.io/files/0vv8moc6/biopharn/5243bb68565995a5950dd31a03f371002b7aa8a4.pdf/BP0922_ezine%20(Watermark)_Linked.pdf [Accessed 15 Jul 2023].

Gibbons, L.A., Rafferty, C., Robinson, K., Abad, M., Maslanka, F., Le, N., Mo, J., Clark, K., Madden, F., Hayes, R., McCarthy B., Rode, C., O'Mahony, J., Rea, R. and Harnett, C.O. (2021) 'Raman based chemometric model development for glycation and glycosylation real time monitoring in a manufacturing scale CHO cell bioreactor process,' *Biotechnology Progress*, 38(2). doi: 10.1002/btpr.3223.

Kirdar, A.O., Conner, J.S., Baclaski, J. and Rathore, A.S. (2008) 'Application of multivariate analysis toward biotech processes: Case study of a cell-culture unit operation,' *Biotechnology Progress*, 23(1), pp. 61–67. doi: 10.1021/bp060377u.

Korstange, J. (2021) Partial Least Squares. Available at: https://towardsdatascience.com/partial-least-squares-f4e6714452aowards Data Science [Accessed 25 Jul 2023]

Rafieyan, S., Vasheghani-Farahani, E., Baheiraei, N. and Keshavarz, H. (2023) 'MLATE: Machine learning for predicting cell behavior on cardiac tissue engineering scaffolds,' *Computers in Biology and Medicine*, 158, p. 106804. doi: 10.1016/j.compbiomed.2023.106804.

Rashedi, M., Khodabandehlou, H., Demers, M., Wang, T. and Garvin, C. (2022) 'Model Predictive Controller Design For Bioprocesses Based On Machine Learning Algorithms,' *IFAC-PapersOnLine*, 55(7), pp. 45–50. doi: 10.1016/j.ifacol.2022.07.420.

Sartorious (2023a) *Biostat® B-DCU*. Available at: https://www.sartorius.com/en/products/fermentation-bioreactors/benchtop-bioreactors/biostat-b-dcu [Accessed 08 Oct 2023]

Sartorious (2023b) *Simplifying Process: BioPAT® MFCS*. Available at: https://www.sartorius.com/download/518648/biopat-mfcs-brochure-en-b-sbi1519-sartorius-data.pdf [Accessed 08 Oct 2023]

Sartorious (2022) *Implementation Guide: Simca®-Online 17*. Available at: https://www.sartorius.com/download/545254/implementation-guide-simca-online-en-b-00039-sartorius-pdf-data.pdf [Accessed 08 Oct 2023]

Stephenson, M. and Grayson, W. (2018) 'Recent advances in bioreactors for cell-based therapies', *F1000Research*, 7. doi: 10.12688/f1000research.12533.1.

Tulsyan, A., Garvin, C. and Ündey, C. (2018) 'Advances in industrial biopharmaceutical batch process monitoring: Machine-learning methods for small data problems,' *Biotechnology and Bioengineering*, 115(8), pp. 1915–1924. doi: 10.1002/bit.26605.

Wold, S., Sjöström, M. and Eriksson, L. (2001) 'PLS-regression: A basic tool of chemometrics,' *Chemometrics and Intelligent Laboratory Systems*, 58(2), pp. 109–130. doi: 10.1016/s0169-7439(01)00155-1.

Yamamoto, T., Taylor, J.N., Koseki, S. and Koyama, K. (2023) 'Prediction of growth/no growth status of previously unseen bacterial strain using Raman spectroscopy and machine learning,' *LWT*, 174, p. 114449. doi: 10.1016/j.lwt.2023.114449.

Appendix

Below are the process parameters used in the research.

Table 2: Process variables/parameters

Parameter Name	Description
Batch	Batch ID
Observation	Date and time observed for each parameter value
AIRSP Value ccm	Air Sparge value
AIRSP ST PT ccm	Air Sparge set point
BASESUB ST PT %	Base pump set point
BASESUB Value %	Indicator of base pump activation from Sartorius MFCS
BASET Value ml	Totalizer value indicative of total base additions from Sartorius MFCS
CO2SP Value ccm	Value for CO2 gas flow from Sartorius MFCS
CO2SP ST PT ccm	Set point for CO2 gas flow from Sartorius MFCS
JTEMP Value °C	Value for bioreactor jacket temperature
JTEMP ST PT °C	Set point for bioreactor jacket temperature
O2SP Value ccm	Value for bioreactor O2 gas flow
O2SP ST PT ccm	Set point for bioreactor O2 gas flow
pH st pt	Set point for bioreactor pH
pH Value	Value for bioreactor pH
pO2 ST PT % sat	Set point for bioreactor pO2
pO2 Value % sat	Bioreactor pO2 value
STIRR Value rpm	Bioreactor agitator value
STIRR ST PT rpm	Bioreactor agitator set point
SUBS A ST PT %	Pump A set point
SUBS A Value %	Pump A value
SUBS B ST PT %	Pump B set point
SUBS B Value %	Pump B value
TEMP ST PT °C	Bioreactor temperature set point
TEMP Value °C	Temperature value inside bioreactor
Time days	Time in days (Parameter created in SIMCA to represent time in days)
AIPSP st pt_value difference	Difference between set point and value for air sparger
BASESUB st pt_value difference	Difference between set point and value for base pump %
CO2SP st pt_value difference	Difference between set point and value for CO2
JTEMP st pt_value difference	Difference between set point and value for jacket temperature
O2SP st pt_value difference	Difference between set point and value for inlineO2
pH st pt_value difference	Difference between set point and value for inline pH

pO2 st pt_value difference	Difference between set point and value for inline pO2
STIRR st pt_value difference	Difference between set point and value for bioreactor agitation
SUBSA st pt_value difference	Difference between set point and value for pump A
SUBSB st pt_value difference	Difference between set point and value for pump B
TEMP st pt_value difference	Difference between set point and value for inline temperature
VCD	Viable Cell Density - Cell concentration measured by Raman

Table 3: Batch IDs

Batch ID
1013.20230215_5079460-0002_Qual Run 1_KCDM
1014.20230215_5079460-0002_Qual Run 1_KCDM
1015.20230215_5079460-0002_Qual Run 1_KCDM
1016.20230215_5079460-0002_Qual Run 1_KCDM
1017.20230215_5079460-0002_Qual Run 1_KCDM
1018.20230215_5079460-0002_Qual Run 1_KCDM
1019.20230215_5079460-0002_Qual Run 1_KCDM
1020.20230215_5079460-0002_Qual Run 1_KCDM
1024.20230215_5079460-0002_Qual Run 1_KCDM
1013.20230315_5081143-0002_QualRun2_KCDM
1014.20230315_5081143-0002_QualRun2_KCDM
1015.20230315_5081143-0002_QualRun2_KCDM
1016.20230315_5081143-0002_QualRun2_KCDM
1013.20230419_5083369-0002_QualRun3_KCDM
1014.20230419_5083369-0002_QualRun3_KCDM
1015.20230419_5083369-0002_QualRun3_KCDM
1016.20230419_5083369-0002_QualRun3_KCDM
1014.20230518_5085672_0001_Sat_SubCu_KCDM
1015.20230518_5085672_0001_Sat_SubCu_KCDM
1017.20230601_5086805-0001_Sat_IV_ENG_KCDM
1018.20230601_5086805-0001_Sat_IV_ENG_KCDM
1019.20230601_5086805-0001_Sat_IV_ENG_KCDM
1023.20230601_5086805-0001_Sat_IV_ENG_KCDM
1017.20221003_Sat_SDIV_5074843-0001
1018.20221003_Sat_SDIV_5074843-0001